AN EVALUATION OF THE SPOKEN LANGUAGE SYSTEM INTERFACE FOR THE VOICE-ACTIVATED LOGISTICS ANCHOR DESK

A thesis presented to the Faculty of the U.S. Army Command and General Staff College in partial fulfillment of the requirements for the degree

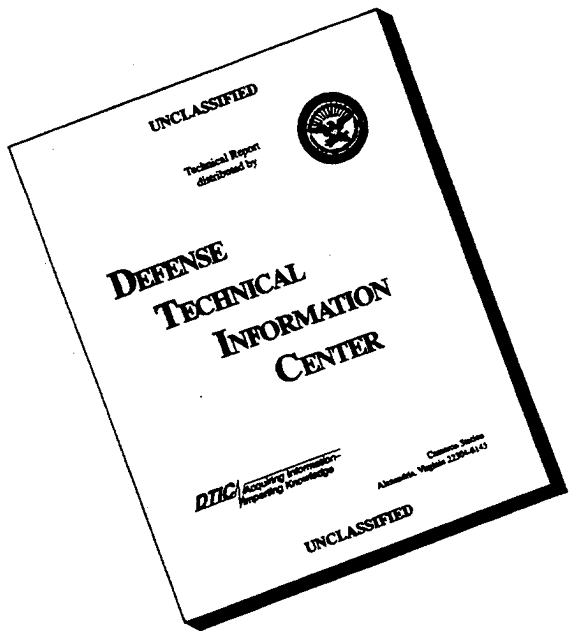
MASTER OF MILITARY ART AND SCIENCE

by

THERON BOWMAN, MAJ, U.S. ARMY
B.A., University of South Florida, Tampa, Florida, 1982

FORT LEAVENWORTH, KANSAS
1996

DISCLAIMER NOTICE



THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
7 June 1996

3. REPORT TYPE AND DATES COVERED
Master's Thesis, 2 Aug 95 - 7 Jun 96

4. TITLE AND SUBTITLE
An Evaluation of the Spoken Language System Interface for the Voice-Activated Logistics Anchor Desk

5. FUNDING NUMBERS

6. AUTHOR(S)
Major Theron Bowman, U.S. Army

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
U.S. Army Command and General Staff College
ATTN: ATZL-SWD-GD
Fort Leavenworth, Kansas 66027-1352

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release, distribution is unlimited.

12b. DISTRIBUTION CODE
A

13. ABSTRACT *(Maximum 200 words)*

This thesis is a preliminary study of the spoken language system (SLS) interface under development for the U.S. Army automated planning and asset visibility system called Logistics Anchor Desk (LAD). The purpose of the study is to determine whether or not there is advantage in the addition of an SLS interface capability to LAD's graphic user interface. One of the uses for automation tools in the military community is the analysis of quantifiable aspects of military operations as input to the military decision-making process. This study looks at the potential impact an SLS interface can have on one of those tools given the current state of the art in speech-based interface technology. Results of the test done with LAD comparing performance with and without the SLS interface indicate several advantages to be gained with the SLS interface. Most notably, participants were able to complete a routine set of tasks in one-third the amount of time. Additionally, participants felt that the system was more usable and would require less user training time. Despite current limitations in the technology, speech-based interface has clear potential for military application.

14. SUBJECT TERMS
Voice-Activated Logistics Anchor Desk, Spoken Language System, Speech Interface, Usability Test

15. NUMBER OF PAGES
87

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

GENERAL INSTRUCTIONS FOR COMPLETING SF 298

The Report Documentation Page (RDP) is used in announcing and cataloging reports. It is important that this information be consistent with the rest of the report, particularly the cover and title page. Instructions for filling in each block of the form follow. It is important to ***stay within the lines*** to meet ***optical scanning requirements***.

Block 1. Agency Use Only *(Leave blank)*.

Block 2. Report Date. Full publication date including day, month, and year, if available (e.g. 1 Jan 88). Must cite at least the year.

Block 3. Type of Report and Dates Covered. State whether report is interim, final, etc. If applicable, enter inclusive report dates (e.g. 10 Jun 87 - 30 Jun 88).

Block 4. Title and Subtitle. A title is taken from the part of the report that provides the most meaningful and complete information. When a report is prepared in more than one volume, repeat the primary title, add volume number, and include subtitle for the specific volume. On classified documents enter the title classification in parentheses.

Block 5. Funding Numbers. To include contract and grant numbers; may include program element number(s), project number(s), task number(s), and work unit number(s). Use the following labels:

C	- Contract	**PR**	- Project
G	- Grant	**TA**	- Task
PE	- Program Element	**WU**	- Work Unit Accession No.

Block 6. Author(s). Name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. If editor or compiler, this should follow the name(s).

Block 7. Performing Organization Name(s) and Address(es). Self-explanatory.

Block 8. Performing Organization Report Number. Enter the unique alphanumeric report number(s) assigned by the organization performing the report.

Block 9. Sponsoring/Monitoring Agency Name(s) and Address(es). Self-explanatory.

Block 10. Sponsoring/Monitoring Agency Report Number. *(If known)*

Block 11. Supplementary Notes. Enter information not included elsewhere such as: Prepared in cooperation with...; Trans. of...; To be published in.... When a report is revised, include a statement whether the new report supersedes or supplements the older report.

Block 12a. Distribution/Availability Statement. Denotes public availability or limitations. Cite any availability to the public. Enter additional limitations or special markings in all capitals (e.g. NOFORN, REL, ITAR).

- **DOD** - See DoDD 5230.24, "Distribution Statements on Technical Documents."
- **DOE** - See authorities.
- **NASA** - See Handbook NHB 2200.2.
- **NTIS** - Leave blank.

Block 12b. Distribution Code.

- **DOD** - Leave blank.
- **DOE** - Enter DOE distribution categories from the Standard Distribution for Unclassified Scientific and Technical Reports.
- **NASA** - Leave blank.
- **NTIS** - Leave blank.

Block 13. Abstract. Include a brief *(Maximum 200 words)* factual summary of the most significant information contained in the report.

Block 14. Subject Terms. Keywords or phrases identifying major subjects in the report.

Block 15. Number of Pages. Enter the total number of pages.

Block 16. Price Code. Enter appropriate price code *(NTIS only)*.

Blocks 17. - 19. Security Classifications. Self-explanatory. Enter U.S. Security Classification in accordance with U.S. Security Regulations (i.e., UNCLASSIFIED). If form contains classified information, stamp classification on the top and bottom of the page.

Block 20. Limitation of Abstract. This block must be completed to assign a limitation to the abstract. Enter either UL (unlimited) or SAR (same as report). An entry in this block is necessary if the abstract is to be limited. If blank, the abstract is assumed to be unlimited.

AN EVALUATION OF THE SPOKEN LANGUAGE SYSTEM INTERFACE FOR THE VOICE-ACTIVATED LOGISTICS ANCHOR DESK

A thesis presented to the Faculty of the U.S. Army Command and General Staff College in partial fulfillment of the requirements for the degree

MASTER OF MILITARY ART AND SCIENCE

by

THERON BOWMAN, MAJ, U.S. ARMY
B.A., University of South Florida, Tampa, Florida, 1982

FORT LEAVENWORTH, KANSAS
1996

MASTER OF MILITARY ART AND SCIENCE

THESIS APPROVAL PAGE

Name of Candidate: MAJ Theron Bowman

Thesis Title: An Evaluation of the Spoken Language System Interface for the Voice-Activated Logistics Anchor Desk

Approved By:

______________________________, Thesis Committee Chairman
LTC Jerry L. Thompson, B.S.

______________________________, Member
MAJ John S. Kane, B.S.

______________________________, Member, Consulting Faculty
LTC James C. McNaughton, Ph.D.

Accepted this 7th day of June 1996 by:

______________________________, Director, Graduate Degree Programs
Philip J. Brookes, Ph.D.

The opinions and conclusions expressed herein are those of the student author and do not necessarily represent the views of the U.S. Army Command and General Staff College or any other governmental agency. (References to this study should include the foregoing statement.)

ABSTRACT

AN EVALUATION OF THE SPOKEN LANGUAGE SYSTEM INTERFACE FOR THE VOICE-ACTIVATED LOGISTICS ANCHOR DESK by MAJ Theron Bowman, USA, 87 pages.

This thesis is a preliminary study of the spoken language system (SLS) interface under development for the U.S. Army automated planning and asset visibility system called Logistics Anchor Desk (LAD). The purpose of the study is to determine whether or not there is advantage in the addition of an SLS interface capability to LAD's graphic user interface.

One of the uses for automation tools in the military community is the analysis of quantifiable aspects of military operations as input to the military decision-making process. This study looks at the potential impact an SLS interface can have on one of those tools given the current state of the art in speech-based interface technology.

Results of the test done with LAD comparing performance with and without the SLS interface indicate several advantages to be gained with the SLS interface. Most notably, participants were able to complete a routine set of tasks in one-third the amount of time. Additionally, participants felt that the system was more usable and would require less user training time. Despite current limitations in the technology, speech-based interface has clear potential for military application.

ACKNOWLEDGEMENTS

I wish to gratefully acknowledge the contibution of Madeleine Bates, Ph.D., to this research project. Although she was not a member of the formal thesis committee, her input was a critical ingredient in my efforts to focus the topic and organize my research. I would have spent valuable time and energy on unnecessary tangents without Dr. Bates'professional advice and kind mentorship.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Purpose

This chapter provides an introduction to the research project. It presents the problem statement and closes with the research question and its importance. To establish the context of this thesis, key definitions of terms and concepts addressed in this paper will be provided as well as background information and relevant historical information. Due to the technical nature of the project, a description of the technology addressed and a review of the current state of that technology is also provided. It includes a description of the limits and delimits on the research question along with the rationale for each and a summary of the research approach.

Problem Statement

A constraint on the effective application of automation for the military decision-making process is the limitations inherent in the current types of man-machine interfaces. An efficient interface, efficient meaning achievement of the maximum human interface bandwidth possible with a given interface model,[1] is dependent primarily on user skills. The current model of human-computer interaction is known as "WIMP" interaction due to its reliance on windows, icons, menus, and pointing realized with a keyboard or pointing device as the interface device.[2] In this interaction model, the bandwidth of data exchange between man and machine is directly related to such user dependent factors as familiarity with software, typing ability, and level of computer literacy. These and other limitations can be decreased somewhat through user training; however, they cannot be completely overcome with traditional man-machine interface capabilities. An ideal interface would be one that requires little or no user training for the interface device and approximates a form of natural human interaction. This type of input device integrated with a graphical user's interface that presents process metaphors (icons) familiar to the average user would immediately decrease the user training requirement and increase productivity.

Speech is the most common form of human interaction. "A speech interface, in a user's own language, is ideal because it is the most natural, flexible, efficient, and economical form of human communication."[3] Current technology does make speech-based interface possible. Although the capability exists, significant limitations remain to a speech-based interface on a computer used to automate tasks in a military decision-making process. The benefits of a speech-based computer interface are enormous. In general terms, it could increase the current gains of automation dramatically when the user dependent limitations of the current man-machine interface were reduced. Until a revolution occurs that drastically alters the current human-computer interaction model, the human limitations inherent in current man-machine interfaces will remain. These limitations can be minimized with a more natural form of interaction with a computer realized by a speech-based interface.

The current state of technology has achieved a significant measure of success with the speech-based interface.[4] The main sponsor of speech understanding research is a Department of Defense agency known as ARPA (Advanced Research Projects Agency). Through its Spoken Language Technology program, research in the field of continuous speech recognition has been encouraged.[5] Recent success in ARPA-sponsored programs indicates the technology is capable of limited military application. This success indicates the potential for integration of a speech-based interface with an automated decision-making tool for military use.

Key Definitions

Before the impact of a technology can be evaluated fairly, it is important to gain at least a basic understanding of it and what it implies. It is also necessary to understand the domain in which the technology is being applied to appreciate the impact realized by the additional capability represented by the technology. Therefore, the following list of terms will help provide an understanding of the context of this thesis.

Efficiency. Efficiency is determined based on three criteria. The first is a comparison of the amount of time required to complete a task using different human-computer interfaces. An interface that can be used to complete a task faster than with another type of interface can be considered more efficient. The second criterion is subjective, in that it requires a comparison of the amount of effort involved in different techniques by a user to complete a task. This effort can be measured in many ways, i.e., number of steps a user must go through, the number of times a user must refer to a manual or seek assistance, the level of physical activity required on the part of the user to name a few. The third and most critical criterion for this research project is user impression. From the standpoint of usability, an interface is truly more efficient only when the user perceives it to be so.

Man-Machine/Human-Computer Interface. A physical device or method employed by a computer user to input data or execute a software function. An example of a man-machine interface would be a keyboard, mouse, or light pen.

Military Decision-Making Process. A systematic approach to decision making consisting of six steps: Step one, Recognize and define problems; Step two, Gather facts and make assumptions to determine the scope of and the solution to problems; Step three, Develop possible solutions; Step four, Analyze each solution; Step five, Compare the outcome of each solution; Step six, Select the best solution available.

Speech/Language Understanding Capability. A software and/or hardware configuration on a computer that translates a user's spoken input through speech recognition and translates it into a meaning representation, and thence into a computer programming language that can be executed by the computer.[6] This capability is an intermediate step in the process that enables a computer to recognize and execute a spoken command. It is an attribute of speech recognition within the scope of this thesis.

Speech Recognition Capability. A software and/or hardware configuration on a computer that translates a user's spoken input into a string of words (possibly including some alternatives).[6]

Speech/Voice/Spoken Language Interface. A method of man-machine interface that involves a system of spoken language.

Background

This study is part of an effort by the Army Research Lab to evaluate the development of an automated tool for use by Army logisticians. Fort Leavenworth is a test bed for this evaluation. The primary contractor for the system under development is BBN Systems and Technologies in Cambridge, Massachusetts. BBN is one of the leading commercial organizations in the field of spoken language systems and has developed systems for commercial applications currently in use.

Along with an understanding of some key definitions, it is useful to be familiar with certain historical aspects of man-machine interface technology on microcomputers in general and spoken language technology specifically. This knowledge is essential for the development of realistic recommendations for the possible application of the technology in the military environment. In addition, an understanding of the process involved in speech understanding systems is necessary for the development of metrics used to evaluate the efficiency and usability of the added capability. For example, completing a task with an automated tool without a speech understanding interface in the same amount of time it takes with a speech understanding interface does not necessarily indicate "no value added." The speech understanding interface may in fact allow more variables to be considered in the same amount of time and thus increase the quality of the completed task.

The current mode of human-computer interface can be traced back to the invention of the typewriter. In June of 1962 when Teletype shipped its first Model 33 keyboard and punched-tape terminal used on many early microcomputers, the design was based upon a typical QWERTY typewriter keyboard designed by Christopher Latham Sholes dating back to 1867. There have been attempts to change the standard keyboard layout but none have been adopted as a new standard. The next significant change in the man-machine interface occurred in May of 1981. The Xerox Corporation unveiled a microcomputer called "Star" that used a mouse and icons in addition to the traditional keyboard for user input. The computer itself failed to become a commercial success, but the input model had a significant impact on the microcomputer market. The next advance in interface capability can be attributed to Apple and its computer called "Lisa." It was the first microcomputer to use a graphical user interface. Although interface hardware has changed very little from the standard keyboard and pointing device, Microsoft's introduction of "Windows" in November of 1983 and its subsequent upgraded versions could be viewed as the last significant improvement in the human-computer interface model.[7]

A quick overview of the history of spoken language interface technology can be seen by tracing the advances in ARPA's Human Language Systems Program. Although industrial research has been conducted outside the ARPA community, ARPA has been involved in this technology for more than twenty years. It has focused its effort on two domains, continuous speech recognition based upon Wall Street Journal text and the development of an Air Travel Information Service that integrates speech understanding capability into an automated airline scheduling service. Despite many years of research, the most significant achievements in this area have only occurred in the last four years. Spoken input as a means of human-computer interface is just beginning to become practical. "Current speech recognition systems still fall short of human capabilities of continuous speech recognition,"[8] but the technology has advanced to a state where some limited tasks can be performed. Despite the current limitations of a speech-based interface, the computer industry is proceeding on the assumption that speech is the next major component in the human-computer interaction model.[9]

There are many different types of speech understanding systems in use today. These systems can be categorized as either speaker dependent or speaker independent. A speaker dependent system must be programmed or "trained" by a specific individual for use by that individual. A speaker independent system can be used by several individuals without being "trained." The system manufacturer will program a predetermined vocabulary into the system that can be recognized when spoke by a variety of users. A variation of these is a speaker adaptive system that operates as a speaker independent system but can become tuned to the specific voice of individual users.[10]

Both of these primary types of speech understanding systems operate in one of three modes: discrete, connected, or continuous. The discrete mode requires a user to isolate each individual word in a phrase by a short silence. These types of systems are easier than others to implement because the exact extent of each word is known and therefore easier to decode by the system. Although it is easier for the computer, it is not natural or "convenient" for the user. Connected speech involves a more natural interface for the user but involves longer delays for the computer than a true continuous speech system. In connected speech, the spoken words of a user go into a buffer memory and then are presented to a processor during pauses by the user. The continuous speech system eliminates the buffer, continuously recognizes what is being said, then takes appropriate action. Continuous speech systems present the most natural interface for a user but represent the most advanced aspects of the technology.[11]

Each of these systems provides numerous advantages at present. Within the civilian community, speech understanding systems are being applied in many practical ways. Speech provides a shortcut for some basic tasks that require several steps with a keyboard or pointing device. For example, opening a file, changing a font, changing a drawing tool, or conducting constraint-based information retrieval ("find all the E-mail messages from Nancy received after October") can be accomplished. There are systems available currently that use speech understanding systems in conjunction with a word processor, data base, or spreadsheet to generate free text at up to 25 words per minute. Many telephone companies have implemented automated operator-assisted telephone call services with speaker independent, isolated word speech understanding systems. Research is currently underway in the area of home automation. Speech understanding could be used to control climate, security systems, appliances, entertainment systems, lights, and various other elements of the home environment. This is certain to have a positive impact for the average person. The potential impact this technology could have on the disabled is tremendous. Its immediate effect would be a greater number of disabled individuals who could experience a higher degree of self-sufficiency.

Within the military community, the application of speech technology is being researched and demonstrated in several different areas. The military aviation community has done a significant amount of research on the advantages of speech-based interface capabilities to aid in cockpit management. In the late 1980's, the U.S. Army Avionics Research and Development Activity along with the Aviation Test Board conducted tests with a JOH-58C observation helicopter equipped with a continuous speech system. The system was used in lieu of manual switching for navigation systems, communications systems, and the Airborne Target Handover System. "The study concluded that voice technology has matured to the point where it can be used as an alternative to manual input methods."[12] The U.S. Air Force has conducted research in the same area as part of their Advanced Fighter Technology Integration F-16 aircraft. In addition, France and the United Kingdom have been doing research on a variety of platforms to integrate a speech understanding system into cockpit design to assist in cockpit management. Each of these programs used the system for a variety of tasks which included communications control, navigation system management, flight data display management, weapons system release parameters, and commanding autopilot systems. In each of these tests, speech understanding systems were used successfully when a very small, constrained vocabulary was used.[13]

Given the current state of speech understanding technology, what is the difficulty involved in developing a computer that recognizes natural human speech? Primarily, it is the variables that exist in spoken language communication that cause the difficulty. First of all, every voice is different and reflects a distinct sociolinguistic background. Regional accents and intonations, gender, age, and other factors create a wide variety of possible signal characteristics associated with a voice input. Another random variable is external factors such as background noise and microphone quality that affect the input signal. Potential vocabulary presents a significant problem in that there are a variety of words and sentence structures that can convey the same underlying message. What also must be considered is that voice systems require significant processing resources and must be used with very advanced and sophisticated hardware in most cases. Finally, spoken language communication is an active process that requires a wide base of knowledge of the talker and the listener. This wide base of knowledge from multiple sources that provide such knowledge as syntactic and semantic framework or discourse context is difficult, to say the least, to reproduce in an automated process. Despite these hurdles, recent achievements in speech technology are promising.

Description of the Voice-Activated Logistics Anchor Desk Speech Understanding Interface

During Desert Shield/Desert Storm, there was little asset visibility of supplies en route to the theater of operation. This lack of visibility added to the difficulty of logistic planning and execution for a theater level operation. With the introduction of split-based operations in Army doctrine and the increased prospect of providing logistic support to multiple theaters for operations other than war, the need for a distribution system with total asset visibility at all levels has become apparent. The Logistics Anchor Desk (LAD) was developed as a tool to maximize the advantages gained in total asset visibility initiatives.

The Logistics Anchor Desk is a microcomputer that interconnects to worldwide logistics data bases to provide situational awareness and knowledge-based decision support tools to rapidly plan and analyze logistics support actions that support the commander's intent.

> With LAD, planners can determine the required equipment and personnel densities, identify support for mission critical units or items, provide projections and summaries of sustainment issues, "look ahead" to forecast densities and stocks, monitor the performance of logistical systems and units, and monitor ongoing deployment actions along with the status of critical items.[14]

The basic components of the speech interface for the Voice-Activated Logistics Anchor Desk (VALAD) are common among typical speech understanding systems. The BBN version of the spoken language system is called "Hark" and consists of a speech recognition system and a natural language understanding system. The speech recognition system matches the acoustic signal of the spoken words against stored sequences of sounds to determine which words were spoken by the user.

The natural language understanding system takes those words and produces a "meaning structure" that represents the literal meaning of the utterance from the user. A data base interface then translates the meaning into a computer language command and retrieves the answer to the user's question. [15] The uniqueness of BBN's Hark system is that it operates on commercial, off-the-shelf, Unix based hardware. In addition, Hark was designed with the hidden Markov model approach to continuous speech recognition, a technique that has shown itself to be superior to other approaches. [16]

Limits/Delimits on the Topic

The topic of this research project is limited to a specific system, the Voice-Activated Logistics Anchor Desk. To further limit the scope, an evaluation of the utility of that system for tasks that would be performed by a logistician assigned to an Army corps headquarters staff will be conducted. To help focus the effort further, the research effort will be limited to the fuel (Class III) and ammunition (Class V) requirements involved in the deliberate planning process of a corps planner. To delimit the research effort, data greater than ten years old will not be used. In addition, data on speech-based systems that do not use the same basic process and represent the same basic capability as the VALAD speech interface will not be used. Issues involving projected costs of the system or reliability, availability, and maintainability will not be addressed.

The Research Question and Its Importance

Given the demonstrated success of speech-based interface technology and its potential impact on automated tools used for military decision making, the following research question is posed: "Is a spoken language interface as part of the graphical user interface for the Logistics Anchor Desk more efficient than the graphical user interface alone?" The scope of this question is defined by the tasks performed by a logistics planner at an Army corps staff level. To establish some benchmark to measure efficiency and determine whether or not the spoken language interface increases overall efficiency, a secondary question arises. That question would be, "What specific tasks are performed by a logistician that can be performed by the Logistics Anchor Desk?" To determine the tasks that could be performed by the system, it is necessary to determine the current capabilities and limitations of the system and those of the spoken language interface, hence, the basis for a third question. There are tasks that the Logistics Anchor Desk is capable of performing that cannot be performed with the speech understanding interface due to a limitation of the technology in its current state. Before overall efficiency can be compared between the two configurations of the system, it is necessary to determine which tasks can be performed with or without the speech interface. Once these tasks are evaluated, those results can be viewed in light of the total system capability and some measure of efficiency can be determined. The determination of efficiency will be based as much on quantitative data from an evaluation as qualitative data from comments of participants and observers in the evaluation process. A more detailed description of criteria that will be used is provided in chapter 3.

The military decision-making process fosters effective analysis of a situation by enhancing the application of professional knowledge, logic, and judgment. The Army views the military decision making process as the product of two disciplines, science and art. Many aspects of all military operations are quantifiable components. This is the "science" aspect of the process. Others are not so concrete and fall under the heading of "art." The military commander as the primary decision maker is constantly challenged to arrive at the best possible decision in the available time through thorough and unemotional analysis of quantifiable components, available facts and assumptions. His analysis of the "science" aspect coupled with his experience, or his application of the "art" aspect, is the basis for a decision.[18] These two components of the decision making process are known as command and control. "Command is the art of war within the domain of the commander; control, as the science of war, is within the purview of the staff."[19] It is the control aspect of the decision making process that can be most affected by automated tools and, therefore, have a positive influence on the command aspect. Of the available time a commander has to make a decision, the largest portion is dedicated to the control aspect. This component includes gathering facts, making assumptions, computing requirements, projecting change, and analyzing performance, to name a few. The control function involves the manipulation of available data to present a current or projected scenario. Projections are based upon historical experience under similar scenarios that has been analyzed and canonized for future reference. The application of automation within the military in this phase of the decision making cycle has had a positive impact thus far on the control function and provides a commander with more possible solutions given the same amount of available time. It is, in fact, an expressed goal of the Army's Force XXI doctrine for the future to digitize the battlefield for the commander and enable a "virtual presence" that fosters better decisions. Currently, though, there is room for improvement.

As with other automated functions, the contribution an automated system makes is a function of the number and abilities of the trained users. Once again, the critical node becomes the man-machine interface. If the interface is easy to use, more people will be prone to use the automated system. Conversely, if the interface requires significant training time and practice ("significant" would be user defined) for the development of skills required to take full advantage of the automated system involved, fewer users will make the effort or "can afford" to invest the time. Therefore, the more natural or graceful the interface, the greater the number of users available to operate the system. A speech understanding interface on an automated decision-making tool would provide such a graceful interface for users. Until a paradigm shift occurs away from the current human-computer interaction model which relies heavily on the graphical user interface, the addition of a speech understanding interface with the graphical user interface provides the greatest promise of increased efficiency.

Research Design

The research design for this thesis involves three types of research methods. First, only research material reflecting the limits and delimits established will be used. The second method is to conduct interviews of logisticians assigned to Fort Leavenworth to develop an answer to the secondary research question. The third method is to conduct a system evaluation with a focus on usability.

The combination of all three of these research methods will provide the framework to formulate a conclusion as to the utility of the Voice-Activated Logistics Anchor Desk at its current level of development and identify improvements that could be made to that system to increase its usability. Conclusions will include comments specific to that system and in general concerning the current state of speech interface technology and its application to automated decision-making tools in the military.

Endnotes

[1]Joel N. Orr, "In Pursuit of Immediacy," Computer-Aided Engineering 14 (February 1995): 66.

[2](1995, November). "Human-Computer Interaction Program." [On-line]. Available : http://www.csto.arpa. mil/ResearchAreas/Intelligent_Systems/Human_Computer_Interaction.html.

[3]Victor Zue, "Research Overview," Laboratory for Computer Science, Massachusetts Institute of Technology, Annual Research Summary: Spoken Language Systems (Cambridge: MIT Laboratory for Computer Science, 1995), 3.

[4]John Makhoul, Fred Jelinek, Larry Rabiner, Clifford Weinstein, and Victor Zue, Spoken Language Systems (Cambridge: BBN Systems and Technologies, 1990), 483.

[5]Alexander I. Rudnicky, Alexander G. Hauptmann, and Kai-Fu Lee, "Survey of Current Speech Technology," Communications of the ACM 37 (March 1994): 53.

[7]Ibid.

[6]Dr. Madeleine Bates, E-mail to LTC Larry Ground, Subject: FW:VALAD Research Prospectus for MMAS, Fort Leavenworth, Kansas, September 15, 1995, 4:23PM.

[8]Ken Polsson (4 November, 1995). Chronology of Events in the History of Microcomputers [On-line]. Available: http://www.islandnet.com/~kpolson/comhist.htm

[9]Zue, 3.

[10]Rudnicky, 52.

[11]Ibid., 54.

[12]Mary L. Hoeller and Claudia R. Salerni, "Using Continuous Voice Recognition to Operate the Research Evaluation System Analysis (RESA) Wargame" (Master's Thesis, Naval Postgraduate School, 1990), 4-6.

[13]Dominick Pondaco, "Voice Recognition in Army Helicopter Operational Environments," in Proceedings of the 1989 AVIOS Voice I/O Systems Applications Conference (September 1989), quoted in Mary L. Hoeller and Claudia R. Salerni, "Using Continuous Voice Recognition to Operate the Research Evaluation System Analysis (RESA) Wargame" (Master's Thesis, Naval Postgraduate School, 1990), 9.

[14]C. J. Weinstein, Opportunities for Advanced Speech Processing in Military Computer-based Systems (Boston, MA: Massachusetts Institute of Technology, Lexington Lincoln Lab, 1991), 11.

[15]BBN Systems and Technologies, Logistics Anchor Desk User's Manual, BBN Report No. 8074 (Cambridge: BBN System and Technologies, 1995), 1.

[16]Madeline Bates, Robert Bobrow, Francis Kubala, Robert Ingria, John Makhoul, Scott Miller, Long Nguyen, Sandra Peters, Richard Schwartz, David Stallard, and George Zavaliagkos, Usable, Real-Time, Interactive Spoken Language Systems (Cambridge: BBN Systems and Technologies, 1994), 22.

[17]John Makhoul and Richard Schwartz, State of the Art in Continuous Speech Recognition (Cambridge: BBN Systems and Technologies, 1995), 9959.

[18]Department of the Army, US Army Command and General Staff College, ST 101-5, Command and Staff Decision Processes (Fort Leavenworth: USACGSC, February 1995), 1-2.

[19]Ibid., 1-1.

CHAPTER 2

LITERATURE REVIEW

Purpose

This chapter provides a review of literature used during this research project. This review is limited to those items that most directly apply to the primary and secondary research questions, as well as key resource material. The intent is to demonstrate the validity of the problem statement and support the primary research question. In addition, this review will provide the reader with information used and referred to throughout this work. This material provides the basic foundation for an understanding of the problem and the perspective from which recommendations will be made.

Overview of Reference Material by Type

Literature used for this research project can be categorized into five main types: books, articles, unpublished materials, technical reports and reviews, and the Internet. These types represent commercial, government, and academic material. They also involve the application of, and research underway in, speech interface technology. Within each type, material presented tended to represent a single aspect of the topic. For example, all of the technical reports dealt with the research or evaluation of some aspect of speech interface technology. There is clearly an overwhelming amount of information available that impacts specifically on the problem statement. Most of the literature available was in the form of reports, reviews and unpublished material. There seems to be few books published that deal with speech interface technology specifically. The books used dealt with automated systems in general, not any specific interface capability. It is only within recent years that this technology has evolved sufficiently for application, this might explain the difficulty in locating books that deal with this specific technology.

There are numerous articles in a wide range of magazines and journals that deal with the application of speech interface technology. These sources ranged from commercial to academic to professional. This fact alone is a good indicator of the potential impact this technology can have in our culture and provides a good measure of the current state of the technology. There were several areas of commonality between the articles, both in the type of tasks performed and in limitations recognized. This area of the research helped form a level of expectation of the performance of the speech interface for the Voice-Activated Logistics Anchor Desk. This is an important aspect of the research because after the initial cursory review of material available on the subject, the author's level of expectation was much higher than it was after a more detailed review of the technical aspects of the technology. This is not mean to detract from the tremendous potential present, this is a very promising development in computer science. It is, however, that the techniques used to apply speech understanding represent a synergistic effect of several disciplines within computer science and phonetics. The variables that exist when dealing with human speech represent the greatest challenge to widespread application of speech understanding technology at this stage of its development.

There is a voluminous amount of unpublished material available on this subject. Even after limiting the research effort to material dealing with military applications of the technology, there is an ample supply of information. This source of information was useful for both background information and for examples to use when forming a research design. A majority of the unpublished material came from theses written at the Naval Postgraduate School in Monterey, California. Based upon the information available, it appears that the Navy is experimenting with a greater variety of applications of this technology than the other services. The Air Force and Army have done a fair amount of research, but it seems to be focused primarily on application of speech understanding in airplane and helicopter cockpits to reduce pilot workloads.

The technical reports and reviews provided the best source of information on the mechanics of speech understanding. They represented a wide variety of applications for the commercial and military communities. The detailed nature of these sources provides an insight into the challenges involved in this technology. As stated earlier, the common challenge is the variables present in human speech, but there are still others to be considered. Other variables involved in the application of speech understanding include the presence of background noise, the varying quality of microphones and other hardware, processing capability of the computer involved, and developing a structural framework and parameters for natural spontaneous human speech to name a few. Insights gained from a review of these sources of information help develop a realistic perspective of the current state of speech understanding technology when forming recommendations about a specific application such as the Voice-Activated Logistics Anchor Desk.

The Internet has provided a wealth of information for research. The primary focus on Internet sources is academic and government institutions that are conducting or sponsoring research on natural language processing systems. The primary government source, ARPA, is extremely useful as a springboard for other sources. The ARPA site contains a list of research projects sponsored by that office as well as technical reports and background information on the history of technologies within the scope of ARPA's interest. Several universities worldwide are conducting research with natural language systems. The Massachusetts Institute of Technology, Cambridge University in England, the University of Edinburgh in Scotland, the Oregon Graduate Institute, and the University of Rochester were all excellent sources of information. Each of these institutions offers advanced degree programs associated with natural language processing systems and is conducting research in the field.

Review of Key Source Material

Although few books deal exclusively with speech understanding technology, several provide a methodology for evaluating automated systems in general. The most significant of these for this study is entitled *A Practical Guide to Usability Testing* by Joseph S. Dumas and Janice C. Redish.[1] While not providing specific information concerning speech interface technology, this book provides a unique and very appropriate methodology for evaluating a speech understanding interface. The importance of this methodology is that its focus is on the usability of a product rather than its functionality. Functionality is a critical aspect of performance, but it is not sufficient in and of itself. The key to the success of a product, in this case a speech understanding interface, is usability. Functionality refers to what a product can do, whereas usability refers to how people work with the product. As discussed earlier in chapter 1, an automated tool may have the capability to increase productivity, but if a user is not convinced that the investment of time to learn how to use it is worth the benefit gained, the automated tool becomes essentially useless. Here in lies the concept of usability. Usability's focus is on the user, not the product. "People's tolerance for time spent learning and using tools is very low."[2] For this reason, usability is a critical concern in the commercial market because of its impact on sales. The implication for military application is that the resources expended for the development and fielding of an automated tool will be wasted if the tool is not usable to the soldier in the field. A system like the Voice-Activated Logistics Anchor Desk may very well have the potential to significantly increase the productivity of a staff member involved in the decision-making process, but the staff member will not use the tool if he or she is uncomfortable with it or "cannot afford the time" to learn how to use it. It is ultimately the user that determines whether or not a tool is easy to use. When a tool is so "functional" that it becomes a challenge within itself to learn how to use rather than a less complicated means to an end, a user is less likely to use the tool. A good example of this struggle between functionality and usability is programming a VCR to tape a television program. The goal is not to operate the VCR, it is to tape the show to watch later. To some people, the goal of learning to program the VCR is more difficult to achieve than taping the show by some other method.

The idea of usability is the key to answering the primary research question, "Is a spoken language interface as part of the graphical user's interface for Logistics Anchor Desk more efficient than the graphical user interface alone?" Although this question could be answered based upon the functionality of a spoken language interface, the addition of this interface to the Logistics Anchor Desk would ultimately be useless if soldiers determine it to be unusable. Therefore, the usability testing of a tool must be based upon actual tasks performed by users and the users tolerance for time and effort expended on learning to use the tool. A more detailed description of usability testing will be provided in chapter 3. The major advantage to taking a usability approach of testing the Voice-Activated Logistics Anchor Desk for this research is that it is appropriate at any phase of the design and development process of a product. The goal of a usability test is to uncover problems that can be corrected to improve a product. The alternative to this approach would be a research study. The goal of a research study is to test the existence of a phenomenon, in this case increased efficiency. Because the Voice-Activated Logistics Anchor Desk is still under development, a research study at this point would not result in valid findings. By taking a usability approach, one is able to evaluate the system at this stage of development, uncover existing problems, determine the probability of improvement in those areas uncovered, and project the impact on the efficiency that could be gained with the system when it is fully developed. Another advantage to this approach is that by design, a usability test requires a fewer number of participants selected as a convenience sample rather than a scientific sample for the research study. The same tasks are performed as in the research study, but more weight is given to the qualitative data gathered through observation and comments by the participants than in the research study.

Many articles deal with various aspects of natural language processing systems. Overall the articles provide a good source for understanding the current state of the technology and how it is currently applied. The articles themselves were from a wide range of periodicals and represent a wide spectrum of interests. A majority of the periodicals were computer related, but a good number were not.

One key article was contained in the published proceedings of the 1994 ARPA workshop on human language technology. The article "Towards Better NLP System Evaluation" deals specifically with key elements of an evaluation methodology for natural language processing systems. Although the methodology presented is more from a functional perspective than from usability, the author does provide an insight into the various elements and variables present in a natural language processing system. The author's basic premise is that there is currently no good evaluation methodology for natural language processing systems. She presents a methodology with key elements under the headings of performance factors, evaluation criteria, test data and assessment strategy. The essential problem with evaluating a natural language processing system is that the evaluation could range from a complete "end-to-end" system evaluation for a system devoted to natural language processing to an evaluation of some component of a natural language processor. In between these two extremes lay systems that use a natural language processor as a subsystem, but include other subsystems that are non-natural language processors, e.g., the Voice-Activated Logistics Anchor Desk which uses multiple interface processors. An evaluation of an entire system provides the best functional evaluation of the natural language processing system but it represents the most problematic approach to take. On the other hand, a limited evaluation of the performance of a component of the natural language processing system is less problematic, but is exactly that, a limited evaluation. Regardless of the scope, one variable remains critical to any evaluation, the context surrounding the subject being evaluated. This context is an important variable for a functional test as well as a usability test. For the evaluation of the Voice-Activated Logistics Anchor Desk, a description of the context is provided through the research question and the limits and delimits on the topic. Steps taken to insure the evaluation remains within that context are fully described in chapter 3. "It is clear that there is no single correct way to evaluate an NLP system."[3] What is important is that the evaluation must be as comprehensive and systematic as possible within the context of the subject being evaluated.

When evaluating performance factors it is essential to consider the system involved, in this case the Voice-Activated Logistics Anchor Desk, and the environment in which that system is used because both contain variables that influence overall performance. The tendency is to pay too much attention to the system being evaluated and not to the environment in which it will operate. This will be an important consideration for the evaluation of the Voice-Activated Logistics Anchor Desk.

When determining evaluation criteria, it is useful to categorize performance under effectiveness, efficiency, or acceptability. Performance criteria can then be determined within each of these categories and an appropriate measure and method then applied. It is important to note that each of these criteria can be given either a quantitative or qualitative measure for evaluation. Determining which measure to use can focus the evaluation on usability rather than functionality.

Test data used for the evaluation must meet three important requirements: it must be realistic, representative, and legitimate. Data for the evaluation includes both the evaluation data, i.e., the specific question by a participant of the evaluation to the Voice-Activated Logistics Anchor Desk, and the answer data. Each of these aspects of the test data will be discussed in chapter 3.

The methodology in this article for determining an assessment strategy provides the framework for an effective evaluation regardless of scope. A methodology appropriate for any natural language processing system evaluation can be designed based upon the answers to two sets of questions. The first set is as follows:

1. What is the motivation for the evaluation?
2. What is the specific goal of the evaluation?
3. From which perspective will the evaluation be conducted?
4. What interests are prompting the evaluation?
5. Who is the audience for the evaluations findings?

Having answered this first set of questions, the evaluation design can be formed from a second set of questions as follows:

1. What orientation (e.g., intrinsic or extrinsic) will the evaluation take?
2. What kind of test (e.g., investigation or experiment) will be conducted?
3. What type of evaluation (e.g., "black box," input changes only or "glass box," system setup changes) is it?
4. What form of yardstick will be used?
5. What style of evaluation (e.g., indicative or exhaustive) will be used?
6. What mode (e.g., quantitative or qualitative) will be used?

Another article that serves as a key source of information was one entitled "Survey of Current Speech Technology." The authors provide a good overview of the commercial applications of speech technology, the main challenges to the application of the technology commercially, and current research in the field. There were several points made in this article that apply to this research project. First of all, "computer manufacturers are proceeding on the assumption that speech will become an important component of the computer interface"[4] despite the wide variety of interfaces available on the market and limitations that currently exist in a speech understanding interface. It is industry's view that the speech interface will augment current interface modes rather than replace them in the near future. This is an important point to remember when considering recommendations for improvements in a speech interface for military application. With a large commercial market comes a greater scope of effort for improvement in the technology and a greater possibility that a limitation noted in military applications is likely present in commercial applications and will probably be overcome. This also raises the point that as the popularity of the speech understanding interface grows and becomes more widely used, it will become more familiar and comfortable for users in the work place and require little additional training. Another useful aspect of this article is the brief synopsis it gives of speech understanding technology. It briefly defines the standard dimensions, i.e., processing techniques, found in current speech understanding systems.

A majority of the unpublished materials that have an impact on this project are masters' theses that deal with topics involving a military application for speech understanding technology. The first of these, "Using Continuous Voice Recognition to Operate the Research Evaluation System Analysis (RESA) Wargame," deals with the operation of an automated wargame system used to train naval officers to respond quickly and correctly to a threat scenario. The goal of the study was to evaluate the use of a specific piece of speech recognition hardware integrated with the existing automated system. There are significant similarities between that study and this one. It made conclusions concerning the limitations of the hardware and software used, it identified advantages and disadvantages to the speech interface, and it provides user input concerning the utility of the speech interface. Another thesis "Continuous Speech Recognition as an Input Method for Tactical Command Entry in the SH-60B Helicopter" looks at the utility of a speech recognition interface as an input

method to reduce the task load of an Airborne Tactical Officer on board the Navy SH-60B helicopter. The thesis was experimental based and involved ten subjects entering frequently used commands into an automation tool used by the tactical officer by two methods: voice and keyboard.
Conclusions were drawn concerning the feasibility of incorporating a speech recognition system on an existing system.

One of the theses, "The Telecommunications Emergency Decision Support System as a Crisis Management Decision Support System," involves a survey of emerging technology that can be applied to an existing network of systems known as the Crisis Management Decision Support Systems. Among the technologies surveyed is voice recognition to enhance crisis decision-making support provided by the existing system. There is an obvious connection between the crisis management decision making process discussed in that thesis and the military decision-making process.

In another thesis "Speech Recognition Application in C.I.C." the author conducted experimental-based research on speech recognition for data input to an automated combat information center. Displays in the combat information center provide input to decision makers during the conduct of a military operation. This study provides user input concerning the utility of a speech interface and measures the increase of productivity realized by a more efficient man-machine interface.

Of the available technical reports and reviews, a few are of particular interest. One is a report completed at the Massachusetts Institute of Technology that is based on a study of military applications of advanced speech processing technology. The study was composed of three major elements, all of which apply to the topic of this study. The first element is an overview of current efforts in military applications of speech technology, the second attempts to identify future applications of speech technology within the military. The third element identifies problem areas where research is needed to meet application requirements. There are several military applications reviewed in the first element of this study. The ones that contribute directly to this research deal with speech recognition in fighter aircraft, military helicopters, battle management, and air traffic control training systems. All of these applications involve the graphic display of information requested by a user through a speech interface. In some cases, the systems involved execute tasks based upon voice commands input through a speech interface. In a related study conducted by the Arizona State University, human factors issues associated with voice technology in the Army's Light Helicopter, Experimental (LHX) are summarized. This study also includes an overview of current state-of-the-art speech recognition technology.

Another useful technical report is one completed by BBN Systems and Technologies. This is a final report on an ARPA-sponsored project intended as the next advance in man-machine interaction by developing a high accuracy spoken language system that operates on cost-effective commercial, off-the-shelf hardware. This report provides a detailed description of the spoken language system used for the Voice-Activated Logistics Anchor Desk and provides test results from previous applications. It is useful as background information and also as a benchmark for comparison of performance in later tests and for this evaluation.

One of the more detailed technical reports was one that describes a documented cockpit design effort by Midwest System Research, Inc., conducted over a fifty-two month period. They researched every aspect of potential cockpit design during this study. Of particular interest is the integration of voice recognition technology and generic voice interface workstation development. This aspect of the study provides insight into the limitations of a speech based man-machine interface.

In the area of voice data entry, Chi Systems, Inc., undertook an effort to determine the feasibility of implementing an effective voice data entry capability for a warehouse automation system. This study provides research results on resulting worker productivity and methods used to overcome existing hardware and software constraints. This application of speech interface technology is similar to a function of the Voice-Activated Logistics Anchor Desk system. The Voice-Activated Logistics Anchor Desk can be used to identify inventory locations and status worldwide. Although this represents a much larger scale than the Chi Systems capability, the user interface model is very similar.

A valuable source of information concerning the limitations of current man-machine interfaces is the top ten list of user-hostile interface designs compiled by the Sandia National Labs. The list covers such aspects as physical packaging and design, software, and supporting documentation for interface hardware. With each design flaw covered are possible improvements that can be made to increase usability. This information provides possible directions for potential recommendations for improvements in the Voice-Activated Logistics Anchor Desk.

The Internet has been a very useful source of information for this study. Information gathered from this source has had a direct impact on the study itself or provided leads for other materials that have had an impact on this work. A good example is the Annual Research Summary by MIT. The report is not available on the Internet, but it is mentioned in various documents available in the computer lab's Internet domain and can be ordered via E-mail from the Laboratory for Computer Science at MIT. This report is extremely useful for understanding the mechanics of a spoken language system. It also contains numerous staff and student reports on different aspects of speech understanding technology.

Another useful source of information is the ARPA domain. Once again, this source provides specific information on speech understanding technology background and current research projects as well as leads for additional information both on and off of the Internet.

Finally, the domain for the Oregon Graduate Institute Center for Spoken Language Understanding contains a wide variety of sources for research and application of the technology. This domain contains background and research information as well as numerous hypertext links to other domains like MIT, University of Rochester, and Cambridge University where a wealth of information is available on the subject of spoken language systems. The Internet has been a key source for up to date information on this topic.

Impact of Literature Review on the Research Design

First of all, the background information gained through the review of available literature forms the foundation for every aspect of the research design. Having no previous experience with speech interface technology, a perspective of the technology itself, the problem statement, and a major portion of the analytical framework used for the evaluation are all formed solely from the literature review. Secondly, within this literature lay the key sources that form the building blocks for the evaluation methodology that will be use with the Voice-Activated Logistics Anchor Desk.

Endnotes

[1]Joseph S. Dumas and Janice C. Redish, A Practical Guide to Usability Testing (Nolwood, NJ: Ablex Publishing Corporation, 1993), 5.

[2]Ibid.

[3]Karen Sparck Jones, "Towards Better NLP System Evaluation," in ARPA Workshop on Human Language Technology, March 8-11, 1994, Merrill Lynch Conference Center, Plainsboro, NJ, by Advanced Research Projects Agency, Information Science and Technology Office (Washington: ARPA, 1994), 98-103.

[4]Alexander I. Rudnicky, Alexander G. Hauptmann, and Kai-Fu Lee, "Survey of Current Speech Technology," Communications of the ACM 37 (March 1994): 52-57.

CHAPTER 3

RESEARCH DESIGN

Purpose

This chapter provides a detailed description of the research methodology applied to the problem statement. It also includes an explanation of the development, execution, and assessment strategy for the system evaluation conducted as part of the research methodology.

Research Design Overview

The technique applied to the problem statement involves three types of research methods. The first method is to conduct a literature review. The second method is to conduct interviews of logisticians assigned to Fort Leavenworth. The third method is to conduct a system evaluation with a focus on usability.

The combination of all three of these research methods provides the framework to formulate a conclusion as to the utility of the Voice-Activated Logistics Anchor Desk at its current level of development and, if appropriate, identify improvements that could be made to that system to increase its usability. Conclusions include comments specific to that system and in general concerning the current state of speech interface technology and its application to automated decision making tools in the military.

Literature Review

The goal of the literature review is to draw necessary background information, an adequate understanding of the technical aspects of speech understanding technology, knowledge of other research currently underway and its potential impact on findings, and various other forms of information dealing with the topic. Included in this material are test results from research already conducted by the Army Research Lab and BBN Systems and Technologies, the primary contractor and developer of the Voice Activated Logistics Anchor Desk. This method and its results are discussed in chapter 2 and will not be covered in greater detail in this chapter.

Interviews

The primary goal of the interviews was to gather data from personnel with experience in logistics planning to be used for the development of realistic tasks for the system evaluation. Using the limits established to define the scope of the research, interviews were conducted with logisticians who specialize in fuel and ammunition. The primary question used for the interviews was, "When planning for an operation at corps level, what questions must a logistics planner answer to prepare his or her input into the commander's decision-making process?" The responses to that question provide examples of the kind of data that can be retrieved with the Logistics Anchor Desk. The specific tasks that must be accomplished to retrieve that data were used for the system evaluation. The identification of realistic tasks is a requirement for the usability aspect of the system evaluation methodology presented as part of the next method.[1] A list of common responses from the logisticians interviewed is provided in appendix A.

System Evaluation

Overview

The system evaluation is the primary component of the research design. Analysis of the data gathered during the system evaluation is the primary means of answering the research question, "Is a spoken language interface as part of the graphical user interface for the Logistics Anchor Desk more efficient than the graphical user interface alone?" The framework of the evaluation is established by answering the eleven questions listed in chapter 2 and by using the guidelines for a usability test in *A Practical Guide to Usability Testing*. The major components of the system evaluation are questionnaires, participants, tasks, observations, and assessment strategy.

The questionnaires were completed in three phases; pretest, posttask, and posttest. Data from the questionnaires were used to record qualitative and quantitative information as well as provide background information for the participants to establish groups for the data analysis. The questionnaires will be discussed in more detail later in the chapter.

The participants were students attending the United States Army Command and General Staff School at the Command and General Staff College who are enrolled in a logistics automation elective course. These individuals were selected because they represent Army logisticians and have had a three-hour introduction to the Logistics Anchor Desk.

The tasks that the participants performed during the system evaluation were derived from the information gathered during the interviews and screened using other criteria. These tasks represent realistic tasks that would be performed by a logistics planner using the Logistics Anchor Desk. Task identification and selection methodology is discussed in more detail later in the chapter.

The goal of the observations was to gather qualitative information about the participant's reactions while using the voice interface. Data from observation is an important part of the overall evaluation because it captures spontaneous remarks and participant "body language" that may indicate ease or difficulty with the interface experienced by the participant. This type of qualitative data cannot be captured with a questionnaire. Inferences based upon an observation made during a session were validated with the participant to insure accuracy.

The assessment strategy for the evaluation involves a statistical analysis of the data collected on the questionnaires. Qualitative data were quantified by having participants rank possible responses to questions. Qualitative input that is not ranked, i.e., general comments and observations, were quantified by determining the number of similar comments by the other participants. A more detailed discussion of the assessment strategy is provided later in the chapter.

System Evaluation Framework

Answers to the questions that follow, as presented in the article "Towards Better NLP System Evaluation" referenced in chapter 2, help establish the basic framework for the system evaluation. The eleven questions are broken into two sets. The first five questions help establish the general aim and scope of the evaluation. They provide parameters within which a more specific evaluation design can be established. Based upon the general aim and scope established with the first set of questions, the second set of six questions help determine the evaluation design. The design includes performance factors, criteria to evaluate the performance factors, test data used for the system evaluation, and the assessment strategy for the data collected.

After the evaluation design was established, guidelines for usability testing were followed to further focus the evaluation. Because the spoken language interface being evaluated is in an early stage of development and due to the limited amount of time afforded in this degree program, a thorough functional test could not be conducted. An evaluation aimed primarily at usability is appropriate for this research project because it is a valid test methodology at any phase of the design and development process.[2] Usability testing focuses on the users and how they work with the product being evaluated. Where a functional test looks at what a product can do, usability looks at the user's determination of the ease or difficulty of use. With this in mind, a qualitative measure to help determine increased efficiency would be the degree to which the participants thought the voice interface was easier to use than the mouse and keyboard alone.

General Aim and Scope

What is the motivation for the evaluation? The motivation is to establish whether a voice interface is a worthwhile addition to the current Logistics Anchor Desk system configuration.

What is the specific goal of the evaluation? The goal is to establish whether the Voice-Activated Logistics Anchor Desk is a better configuration than the current Logistics Anchor Desk configuration.

From which perspective will the evaluation be conducted? The perspective is task oriented. With the specific goal in mind, performance factors were determined for specific tasks. The performance of each configuration of the system for each task can then be compared. This quantitative comparison along with the qualitative comparison from user input can be used to determine if the voice interface configuration is better.

What interests are prompting the evaluation? The interest behind the evaluation lay with the Army Research Lab and its investment of resources, both dollars and people, for the integration of the voice interface into the Logistics Anchor Desk. For that reason, the findings of this evaluation include a recommendation concerning further effort on this project.

Who is the audience for the evaluation findings? Based upon the stated interest, the Army Research Lab staff is included in the audience. Because this evaluation is part of a thesis presented in partial fulfillment of the requirements for a Master of Military Art and Science degree, the audience includes the faculty of the United States Army Command and General Staff College.

Specific Evaluation Design

What orientation will the evaluation take? Given the stated goal of the evaluation and the focus on usability, the orientation of the evaluation is extrinsic. The determination of whether or not the addition of a voice interface is better than the current configuration is based more on external metrics of user input than from internal metrics of functionality.

What kind of test will be conducted? This is an experimental kind of evaluation because it compares two different configurations of a system under controlled conditions to establish a hypothesis.

What type of evaluation is it? Because the evaluation involves input changes to the two system configurations and then comparing system performance with each input, it is a "black box" approach. A "glass box" approach would involve changes in the system setup for each configuration and using a single input.

What form of yardstick will be used? With usability as the focus of the evaluation, performance and qualitative subjective measures are valid yardsticks. Performance can be measured using time and the number of specific events, e.g., references to task instructions, calls for help, observed frustration, unrecognized words by the computer, etc. Qualitative subjective data can be measured by using questionnaires that use horizontal scales that assign numerical ratings to possible user comments. Additional comments can be quantified by determining how many users made the same type comment.

What style of evaluation will be used? This evaluation is clearly indicative rather than exhaustive due to the time and resource constraints. It should be considered a preliminary study as opposed to a detailed research study.

What mode will be used? As previously indicated, quantitative as well as qualitative criteria were used for the evaluation.

With the parameters established by the answers above, an evaluation design can now be worked out to answer the primary research question. Given those parameters, the evaluation is concerned with the preliminary indications of a voice interface's potential improvement of automated tools used by military planners. Indications are based upon descriptive statistics of performance factors for specific tasks and subjective user input. Inferential statistics are used to compare data from the two system configurations used in the evaluation. The statistical analysis will be the basis for a recommendation concerning continued investment into this research.

Usability Guidelines

Because this evaluation's focus is on usability as opposed to functionality, there are specific guidelines unique to a usability study that must be part of the system evaluation framework.[3] Functionality refers to what a product can do, usability refers to how people work with the product. Therefore, usability's focus is on the user more so than the product. The objective of usability is to make the product as easy to use as possible while still performing the function it was designed to perform. "Easy to use" can be characterized by many different factors. First of all, someone may say a system is easy to use if they can do what they need to do with it in a reasonable amount of time. A person may also determine a product is easy to use based upon the number of steps they must perform to complete a task. Still another measure some might use in determining whether or not a system is easy to use is the success they have in predicting the right action to take to complete a task. These and many other characteristics of a product would be determining factors in a user's opinion of a product's usability. Whatever the metric used, it is the perception of the user that ultimately determines whether or not a product is easy to use.[4]

The first requirement of a usability test is that the people who participate in the test must represent real users. It is important to eliminate any bias that may arise in the evaluation if the users are associated with the system in some way, e.g., programmers, sales representatives, design engineers. Personnel who participate in a usability test are called "participants" to help reduce the stress of the testing environment by eliminating the notion that they are being evaluated along with the system. It is critical in a usability test that participants perform real tasks, tasks that they would actually perform with the system if they were using it to do real work. It is also important that observers record what participants do and say during the test. During the actual conduct of the test, a participant's reaction may be an indicator of a problem area. After the test is completed, an observer can ask the participant about their reactions and comments to validate what might be inferred by their reaction and record that as input.

Selection of quantitative evaluation criteria for a usability test is different from other types of tests in that the goal is not to have the product pass the criteria, it is to develop a product people want to use. When using time as a metric, for example, it is important to determine the amount of time that a user would consider acceptable and use that as a baseline measurement along with qualitative input from participants of the test. Something else that must be considered is that those tasks that do use time as a criterion may have different baselines because a user would realistically expect certain tasks to take significantly longer than others and still be happy with its performance.

In some cases, the system response time cannot be improved so it is important to look at other steps that can be taken to reduce user frustration. In addition, time may not always be an appropriate performance criterion for each task. The bottom line is that whatever criteria are selected and whatever metric is used for each criterion, they must reflect what will make the user productive and happy with the product.

As with any other form of test with the goal of product improvement, data must be analyzed. Usability testing is an empirical method that relies as much on quantitative data analysis as it does qualitative data analysis. It is similar to a research study, in that it involves the observation of actual behavior.[5] It takes place in a laboratory environment and uses sample participants from a population of users. Steps are taken to control variables that would make results difficult to interpret. For example, it would be reasonable to develop a task scenario and specific tasks for each participant to perform to reduce the amount unexpected inputs to the system. Objective and subjective measures are recorded in much the same way as a research study.

Although there are similarities between the two techniques, they differ in several ways. The goal of a research study is to test the existence of a phenomenon. Usability testing uncovers problems, not demonstrates their existence. A usability test takes fewer participants than a research study because a large sample is not required. The sample of the population used is only a convenience sample, not a random, scientific sample like that required for a research study. A research study attempts to isolate one independent variable to look at its effect on other variables. A usability test looks at the interaction of all variables and uses observations, participants' comments, quantitative data, and expert knowledge to identify problems. Finally, the use of inferential statistics is at the heart of the analysis and reported data from a research study, they are seldom appropriate for a usability test. Normally, descriptive statistics like, mean, median, ranges, and frequencies are sufficient to identify trends that indicate a potential problem.[6]

Questionnaires

Questionnaires are useful tools for data collection and are appropriate for the technique selected for this evaluation. Three different questionnaires were used in this research methodology; a pretest questionnaire, a posttask questionnaire incorporated in a task scenario sheet to record elapsed time, and a posttest questionnaire.[7] Samples of the questionnaires and the task scenario sheet are provided in appendix B.

Pretest Questionnaire

The purpose of the pretest questionnaire was to collect background information on potential test participants to help in the interpretation of the data collected. Each question has a specific purpose. The first three questions are to determine the participant's level of computer experience. The results of these questions can be used to establish subgroups within the group of participants based on the level of computer experience to see if it had an impact on their impressions of the Voice-Activated Logistics Anchor Desk. The fourth and fifth questions help to indicate each participant's impression of a speech-based interface before they use the Voice-Activated Logistics Anchor Desk. The two questions together indicate if that impression is based on actual experience or what they may have heard about voice interfaces. The final two questions help to determine the level of logistics experience for each participant and can be used to validate whether or not a participant represents an actual user.

Posttask Questionnaire

The posttask questionnaire was integrated into the task scenario sheet given to each participant. Its purpose was twofold. First, it provided the participant a means of recording the elapsed time for each phase of the task scenario. Second, it allowed them to record the number of times they had to refer to detailed instructions to complete each task. The participants who volunteered for this test received a three-hour introduction and overview of the Voice-Activated Logistics Anchor Desk. Although this is obviously not a sufficient amount of time to consider them as trained users, it represents the only amount of formal training available at this institution on the Voice-Activated Logistics Anchor Desk at this time. The task scenario is broken into two phases. The first phase is a series of five tasks that the participant performed without using the spoken language interface on the Voice-Activated Logistics Anchor Desk. The second phase of the scenario involves the same five tasks only during this phase the participant used the spoken language interface to complete them. The questionnaire required the participant to record the start time and finish time for each phase. These times were used as quantitative data for the evaluation.

Posttest Questionnaire

The purpose of the posttask questionnaire was to obtain an immediate reaction to each participant's experience with the Voice-Activated Logistics Anchor Desk. This reaction compared to that participant's impressions recorded on the pretest questionnaire was used to measure the change, if any, in their perception based upon their experience.

The first two questions help determine if the amount of time it took was more or less than what they expected and if they would consider that amount of time acceptable had they been conducting actual work with the system. The third question asked them to make a judgement as to the ease of use of the system. The fourth question is similar to one on the pretest questionnaire and was intended to indicate the change in their impression of the spoken language interface capability. It also required them to record the reason they had that impression. This information helps identify which specific characteristics they used to determine whether or not they were happy with the system. The final two questions are general in nature and were used as supporting data for recommendations contained in chapter 5. The goal of the posttest questionnaire was to provide the data necessary to determine each participant's perceptions of the effectiveness, efficiency, and acceptability of the interface.

Task Selection

The essential requirement to satisfy when selecting tasks for a usability test is that participants attempt tasks that users will want to do with the system. Therefore, the criteria for selection of tasks are that they be realistic, representative of actual tasks, and legitimate.[8] The list of potential tasks to be used for the evaluation came from two sources. The first is the author's knowledge of Army logistics doctrine gained through attending the course of instruction designated to fulfill the requirements of the Master of Military Arts and Science at the United States Army Command and General Staff College. The second source was the interviews discussed earlier in this chapter. The list of potential tasks was then compared with a list of tasks that could be performed on the Logistics Anchor Desk. Version 3.029 of the Logistics Anchor Desk software was used for this evaluation. Those tasks were then compared to a list of tasks that were taught to the group during a three-hour overview of the system. This step was to insure that each participant had received some level of training on each task they would perform for the system evaluation. The tasks that remained that could be completed with the Voice-Activated Logistics Anchor Desk were used for the selection of the five tasks for the system evaluation in the task scenario. The difficulty of each task was determined by the number of steps required to complete it. A single step is counted when a user must execute a key stroke or move the mouse and/or click a mouse button. Tasks that were selected range in difficulty from four steps to ten steps.

This aspect of the task identification and selection process represents a departure from a pure usability test technique in that tasks should be selected to probe potential usability problems. The author chose to control this variable of the test to narrow its scope and make the data analysis easier because the focus is on the usability of the interface and not on the system overall.[9]

The five tasks selected were placed in a logical sequence based upon the requirement for a logistics planner to conduct terrain analysis and determine what effects it would have on the resupply of ammunition to artillery units near an ammunition supply point. The list of tasks and the scenario for the test were provided to the participants on the form of a task scenario.[10]

Test Site

The site selected for the test was a classroom in the Bell Hall academic complex on Fort Leavenworth, Kansas. The computer system used for the test was one of the systems used for classroom instruction and hands-on work by the students. The area where the test was conducted was not isolated or separated from the rest of the classroom nor was the classroom wholly dedicated for this test. In fact, many of the participants completed the task scenario on one side of the room while instruction was being conducted on the other side. There were no video or audio recording devices used during the test. Although this was not an ideal laboratory setting for a typical usability test, it did approximate the environment that the system would be exposed to if it were being used by a logistics planner in a headquarters type complex and, therefore, added to the realism of the task scenario.

Test Conduct

The test was conducted during the latter part of February and during March of 1996. The pretest questionnaires were completed by all the participants during the first week of January 1996 at the beginning of the semester. The test began with a short introduction by the observer and an overview of the test scenario and purpose of the test. Each participant followed the same sequence of events for the test. The tasks were completed first without the spoken language interface and then with the interface. Participants completed the posttask questionnaire during the test and the posttest questionnaire immediately after the test was completed. The entire process took approximately thirty minutes. The observer and the participant were the only personnel in the vicinity of the computer during the test period.

Assessment Strategy

The overall assessment strategy used for the system evaluation involves two types of measures. The first is system performance based upon the time data recorded by participants on the task scenario sheet. The second is subjective based upon the data taken from the pretest and posttest questionnaires.

Performance Measure

The performance measure was based upon a comparison of the mean time it took participants to complete the task scenario without the spoken language system interface with the mean time with the interface. Conclusions about system performance were not based upon the difference between the two means alone. Consideration was also given to the participants impressions about the amount of time it took them to complete the task scenario with the interface. The mean time for each subgroup established based upon computer experience level could also be used to indicate a base line for the amount of time a user at that experience level would feel productive and be happy with for actual work.

Subjective Measure

Subjective measures can be quantitative or qualitative. To help in the analysis of subjective data, participants were provided a selection of answers to each question on the questionnaires along a horizontal scale. Numbers along the scale corresponding to the answer they choose that best reflects their impression were used to quantify subjective comments for analysis. Participants' subjective input also took the form of spontaneous comments during their execution of the task scenario during the test period. Those comments were recorded as observations and compared with similar comments from other participants to indicate impressions about the system. The numerical results of the questionnaires were averaged for analysis. In addition, ranges, frequencies, and standard deviations for each question were analyzed to identify any trends that would influence possible conclusions.

The conclusions drawn from the analysis of the data were influenced heavily by the results from the questions dealing with the participant's overall impression of the interface on the pretest and posttest questionnaires. For that reason, the use of inferential statistical analysis for that data is intended to show whether or not the results are statistically significant. This is necessary to prove that the results cannot be attributed to chance. This is an important aspect of the assessment strategy because of the usability technique selected for the system evaluation. Ultimately it is the user who determines whether or not the addition of a spoken language system interface makes the Voice-Activated Logistics Anchor Desk more efficient to use.

Endnotes

[1]Joseph S. Dumas and Janice C. Redish, A Practical Guide to Usability Testing (Norwood: Ablex Publishing Corp., 1993), 160.

[2]Ibid., 25-26.

[3]Ibid., 22.

[4]Ibid., 4-5.

[5]Ibid., 35.

[6]Ibid., 35-38.

[7]Ibid., 208-212.

[8]Karen Sparck Jones, "Towards Better NLP System Evaluation," in ARPA Workshop on Human Language Technology, March 8-11, 1994, Merrill Lynch Conference Center, Plainsboro, NJ, by Advanced Research Projects Agency, Information Science and Technology Office (Washington: ARPA, 1994), 100.

[9]Ibid., 98.

[10]Dumas, 171-182.

CHAPTER 4

ANALYSIS

Purpose

The purpose of this chapter is to analyze the input provided by potential and actual test participants on the pretest questionnaire, the data on the task scenario sheets collected during the test, and the input provided by participants on the posttest questionnaires.

Pretest Questionnaire Data

The pretest questionnaire was distributed to fifty-one students enrolled in a course entitled "Logistics Automation." This course is part of the elective curriculum at the United States Army Command and General Staff College. In addition to the students, two current instructors and one former instructor (retired during this school year) completed the questionnaire. A sample of the questionnaire is provided in appendix B. Input provided on the questionnaire was analyzed by descriptive statistics using a numerical value assigned to each possible answer a potential participant could select. The primary purpose of this questionnaire was to gain background information on a large sample of Army logisticians and on participants who would be involved in the test. It also serves to reveal impressions the population may have about current speech-based computer interface capabilities. To validate the test format selected, the first two participants who participated were considered a pilot test.[1] Based upon their input and observed problems with the task scenario, the task scenario sheet, task instructions, and posttest questionnaire, changes were made to each item where necessary. The data from the first two participants has not been included in the analysis of the task scenario sheet and the posttest questionnaire. Their input from the pretest questionnaire has been included. Data from those two participants are included in the test results spreadsheet in appendix

D (Table 1) and are highlighted by an asterisk. Although they are entered in the spreadsheet, those cells are text only entries and were not included in the tabulation of the data except for the pretest questionnaire entries.

Based upon the level of computer experience indicated on the questionnaire, results have been separated by experience group. Group one had greater than ten years of experience, group two had five to ten years of experience, and group three had one to five years of experience. These groups represent all of the completed questionnaires to include the three instructors. They are isolated as subgroups for additional analysis. An additional subgroup includes the data from the questionnaires completed by the actual participants in the test. Twenty of the participants were included in this subgroup. This data has been isolated to verify that the results in fact show that the subgroup was representative of the available sample population.

The results from the entire group indicate that the average experience level was five to ten years. They use a desktop type computer either daily or at least two to three times per week. All but four of the respondents used some sort of graphical user interface when they use the computer. Just over one-half (twenty-nine) of the group had never seen or used a speech-based interface. Fourteen had seen one used, ten had tried one once or twice, and only one uses one regularly. The overall impression of the group about a voice interface capability for computers was mildly positive. The average response was 2.31 with three being no opinion and one being extremely positive. All but two in the group has had logistics experience within a corps. Most of that experience was at brigade level and below. Twenty-eight percent of them have had logistics experience at corps level. The group considered its logistics skills slightly better than average. This could be considered about right or slightly under rated because, as a rule, only the top one-half of a peer group is selected for attendance at this institution.

A comparison of the results from the three primary groups and the participant subgroup against the entire sample population indicates that there were no significant differences in the results. Other than the indication that the group with the least amount of computer experience tended to use a desktop computer slightly less, there was no significant difference in the averages among the groups in all other categories. The instructors used a desktop computer on a daily basis and had tried the Voice-Activated Logistics Anchor Desk once or twice, but beyond that they tended to be similar to the rest of the sample population in other categories.

Task Scenario Sheet Data

Data collected on the task scenario sheet was for two purposes. First, the time entries indicate the amount of time it took each participant to complete the task scenario without and with the spoken language interface. Second, it indicates the amount of outside help they required to complete the task scenario. Help was available from two sources.

The first source was a detailed narrative type instruction sheet that describes step by step the procedure necessary to complete each task in the task scenario. A copy of the instruction sheet is provided in appendix C. The steps provided in the instruction sheet represent the fewest number of actions, e.g., key strokes and mouse use, necessary to complete each task. As with most other graphical user interfaces, the Logistics Anchor Desk provides a user with several different options or paths to complete a task. Some options require more actions, ergo more time, than others. To prepare the instruction sheet, a draft user's manual was used to determine the shortest path for each task that was selected for the test. The steps in the instruction sheet represent the technique that a participant could use to complete each task in the least amount of time.

The second source of help was the observer. Most of the instances when assistance was needed from the observer during the first phase of the test were due to a participant departing from or not following the detailed instructions or going from memory and not succeeding in completing the task. During the second phase of the test when participants used the spoken language interface, participants were helped enough to facilitate the discovery of the answer they needed rather than helped in a more direct manner. Each participant was given a four to five-minute overview of the spoken language interface before beginning the second part of the test. This presentation included how to activate the microphone and search the system's dictionary for the word that it would recognize. A demonstration of three similar tasks that they would perform with the system was also provided.

The average time a participant took to complete the task scenario without the spoken language interface was eight minutes and thirty-four seconds. Times ranged from eighteen minutes and forty seconds to four minutes. The group had to refer to the instruction sheet for four out of the five tasks on the average. This average must also be viewed in light of a frequency analysis of the data it includes. Of the twenty participants, sixteen had to refer to the instruction sheet for every task. This would be expected when you consider that it had been two to three weeks since the group received a three-hour overview of the Logistics Anchor Desk. This would seem like an obvious phenomenon given the amount of time since the training occurred. What should be noted however is that each of the instructors that completed the task scenario needed to refer to the instruction sheet for at least one of the tasks. This would indicate that a typical user would probably have to refer to a user's manual on a regular basis when using the Logistics Anchor Desk due to its extensive functionality.

The average time the sample population took to complete the task scenario with the spoken language interface was two minutes and fifty-three seconds. Times ranged from five minutes and thirty-four seconds to one minute and twenty-eight seconds. The distribution for times with the interface is significantly closer to the mean than the distribution was without it. This in itself could indicate some level of increased usability for the typical user over the configuration of the system without the spoken language interface. During this part of the test, four of the twenty participants needed assistance at some point. Assistance was required for no more than two of the tasks for any of the participants. Most of the assistance was to help the participant discover a word in the system's vocabulary to complete a task. In each case, the participant had used a word that was not part of the system's vocabulary. Out-of-vocabulary words were the most common error event during the test. The most common out-of-vocabulary words were added to the system as the test progressed. This is a relatively simple process for a programmer and will be made available in a much simpler form for users in future versions of the software. Times for those four participants are distributed throughout the range of times, therefore, this error did not appear to cause any significant delays.

Group one's average time without the interface was eight minutes and sixteen seconds. This group contained the two times that designate the range for the entire sample population. Of the eight participants in this group, five had to refer to the instruction sheet for every task. Of the three remaining, two were instructors and the third had to use the instruction sheet for three of the tasks. The mean time for this group when they used the spoken language interface was two minutes and twenty-six seconds. Two of the four participants who needed assistance during this part of the test came from this group. Times for this group were slightly faster than the total sample population but were not significantly different.

Group two's mean time without the interface was nine minutes and thirteen seconds. Times ranged from sixteen minutes and five seconds to seven minutes and seventeen seconds. All but one of the participants in this group had to refer to the instruction sheet for every task. The one exception was the instructor included in this group. Their mean time with the interface was three minutes and eight seconds. Times ranged from five minutes and thirty-four seconds to a minute and twenty-eight seconds. This group's time without the interface was clearly slower than the population mean and group one's. Their average time with the interface is not significantly different from the sample populations or group one's.

Although there were only three participants in group three, the mean times for this group fell within one standard deviation of the sample population's mean times. It could therefore be assumed that these times would be indicative of a larger sample. The average time for this group in the first part of the test was ten minutes and eleven seconds. The range was twelve minutes to eight minutes and ten seconds. All three of these participants needed to refer to the instruction sheet to complete the task scenario. Their average time during the second part of the test was four minutes and twenty three seconds. Only one of the three participants in this group required assistance during this phase of the test.

Observations

Observations were recorded for every participant during the conduct of the system evaluation. The purpose for the observations was to record spontaneous comments and body language exhibited by a participant. The observer also attempted to include the number and types of errors that occurred during the test. It was difficult at times during the test to catch all of the errors. Some of the participants would either recover and continue before the observer could recorded the entire event or make multiple errors in rapid succession. In either case, an accurate record was not possible without a software log that recorded each session. Because this is only a preliminary study, the observation record is still accurate enough to identify possible trends.

The observation record indicates that the two largest causes of errors during the second phase of the test were out-of-vocabulary words by the participant and error in word recognition by the system, out-of-vocabulary words being the larger of the two. Most of the recognition errors occurred when participants were executing the task scenario while class was being conducted on the other side of the room. As was true with the out-of-vocabulary word errors, it did not appear that the recognition errors caused any significant delays. Participants recovered very easily for the most part by either repeating the phrase they used or using another word that was easier for the system to recognize. Only six of the twenty participants experienced more than five errors when using the spoken language interface. Of those six, three had distinct voice characteristics. The first one was difficult for the observer to understand because he tended to mumble, the second was from the New York City area and had a very strong, distinctive accent, and the third was a female who spoke with a very soft voice. The female participant completed the task scenario while a class was in session.

General Observations

Most of the participants made some comment about not remembering much from the training they received and went straight to the instruction sheet to begin the test. A few of the participants attempted to execute the tasks by memory. In every case, they either were unable to complete it and had to refer to the instructions or if they did complete the task, it was by some technique other than the one provided on the instruction sheet.

Many of the participants would forget to activate the microphone as they progressed through the second part of the test. After the test when four of them were asked why they seemed to forget that step, every one of them made a similar type comment in that the system was very responsive and they felt so comfortable talking to the computer that they forgot that they needed to activate the microphone before speaking.

The vast majority of the spontaneous comments and facial expressions reflected a positive reaction to the interface. A few of the participants seemed puzzled when an error occurred with the interface but none ever appeared frustrated. The only time participants appeared to get frustrated was during the portion of the test without the spoken language interface.

Posttest Questionnaire Data

The twenty participants for the system evaluation test all completed a posttest questionnaire immediately after the test. The purpose of the questionnaire was twofold. First, it provided a record of their impressions of the system immediately after using it. Secondly, it provided a record of their written comments to include an explanation of their impression of the interface, their recommendation as to whether or not the Army should continue its efforts with the interface, and comments in general about the military application of this technology.

Quantitative Data

The first question deals with the participant's expectation of the performance of the interface. The average response for the entire group indicates that they thought the system was quicker than they expected it to be. The standard deviation of the responses indicates the majority of the group thought that the amount of time it took them to complete the scenario was either about right or quicker than they expected. In fact, only one of the participants had a response outside that range. That participant felt that it took longer than expected. A review of that participant's questionnaire revealed that it was the female with a soft voice mentioned earlier. In her written comments she noted that she felt that she should have spoken more clearly. Her other impressions and written comments would be considered positive or very positive.

The second question asks about the participant's impression concerning the utility of the interface for actual work. The average answer fell between the first two choices, that the interface was as fast as one would like it to be or fast enough for actual work. Two of the participants felt that it was tolerable for some work.

Question three asked them to describe their impression of how easy the interface was to use. Thirteen of the participants thought it was very easy to use, six felt that it was easy to use, and one thought that it was about right.

The fourth question asks the participants to describe their overall impression of the interface. This was the primary question used to determine if there was a significant difference in the participants' impressions of the interface before and after the test. The mean of the answers provided by the twenty participants on the pretest questionnaire was 2.25 on a scale of one to five, one being extremely positive and five being extremely negative. The mean of the answers provided by that same group on the posttest questionnaire was 1.17 on the same scale. Because the standard deviations of the two means overlap, chance cannot be ruled out as the source of the data representing the impression of the participants after using the interface. In other words, the difference between the two means cannot be considered statistically significant. Because there were less than thirty participants in the test, the normal z score method to evaluate the difference between the two means is not considered a good test.[2] A method known as the t test would be appropriate in this case.[3] Based upon the results of a t test evaluation, the probability that the difference between the two means is due to chance is less than 0.005 percent. Therefore, it can be said statistically that the participants' impressions of the interface had increased after using it for the test. In this case, an already positive opinion was strongly reinforced by experience with the system.

The final question asks the participant to describe the nature of their recommendation concerning the Army continuing the integration effort of the spoken language interface into the Logistics Anchor Desk. The possible responses were on a scale of one to five, one being highly positive and five being highly negative. The mean response was highly positive. All of the responses represented the top two choices, fourteen of them indicating their recommendation would be highly positive.

Qualitative Data

Participants were given the opportunity to qualify their response to the last two questions by including written comments on the questionnaire. All of the participants provided comments. Two comments were most common as the reason for the response to the fourth question (How would you describe your impression of the voice interface?). Participants' comments indicated that the two primary reasons their impression was so positive was because they felt the interface was either easy to use or easier than the mouse and keyboard and that it was a time saver. Thirteen of the comments had to do with time savings and ten with ease of use. The next most common answer was that the spoken language interface increases productivity or efficiency. There were five comments along those lines. Other comments were that there would be a reduced training time requirement for the system, it was simple to use or more comfortable compared to the mouse and keyboard method, and that they were impressed at how well the computer recognized their commands. Two comments that appeared only once among the group was that less user errors seemed to occur when using the spoken language interface and that someone was more likely to use the system with the interface installed than without it.

Reasons given for the response to the fifth question were much the same as question number four. The top three comments were that the interface was easy to use, it would reduce the amount of time needed to train an operator, and that it was a time saver. The number of comments in each area was eight, seven, and six respectively. There were three comments concerning the simplicity of the interface and three that indicated it would increase a user's productivity. Two of the participants felt that using the spoken language interface reduced their fear of making a computer error from which it would be difficult to recover. Two others commented that the interface would allow a user to perform other tasks while using the system because it did not require as much attention as the mouse and keyboard did. Other single occurring comments were that more users would use the system because of the interface, that the interface was well suited for military application, and that the interface was less confusing to use than the mouse and keyboard.

The final question on the questionnaire permits participants to make additional comments concerning the interface or comments in general about this capability for military application. Because of the general nature of the question, the spectrum of comments was wide spread. Fourteen of the participants chose to provide additional comments on this portion of the questionnaire. One comment that was common among three of the participants had to do with a concern about the effect of background noise when using the interface. Four other comments appeared at least twice among the group. The first was that the interface should be integrated on other Army systems. The second comment was that the vocabulary of the system should be increased to make it more useful. The third was that this technology had unlimited potential. The fourth comment was that the spoken language interface should be able to perform a compound type command to realize an even greater time savings.

There were ten singular comments. Of those comments, six were general in nature or had to do with the participant's impression and four concerned something specific about the interface. The general comments were that this type of interface made a new computer system easier to use, the technology was impressive, it makes a user more effective, it requires less technical knowledge of the system by the user, it reduces analysis time for a staff planner and therefore the decision time for a commander, and that the Army should continue development of the system. The more specific comments about this spoken language interface were that a user should be able to use both this interface and the mouse/keyboard combination, a user should be able to cancel a command given with the interface by saying something like "cancel," a user should be able to activate the microphone by a voice command in lieu of the mouse or keyboard, and finally that the interface should be in the field now. It should be noted that the version of the spoken language interface used for this test was under development. Many of the changes recommended by the participants are planned upgrades to the system or were available but not active due to instability in the program at this stage.

None of the written comments appeared to be negative or derogatory in nature which reinforces the strongly positive opinion the group had about the interface. Based upon observations, the tone of the written comments, and the fact that many of the participants remained after the test scenario to play with the system, analysis indicates the interface would be readily accepted by the population this group represents.

Endnotes

[1]Joseph S. Dumas and Janice C. Redish, A Practical Guide to Usability Testing (Nolwood: Ablex Publishing Corporation, 1993), 263-269.

[2]Frederick D. Herzon and Michael Hooper, Introduction to Statistics for the Social Sciences (New York: Harper & Row, Publishers, Inc., 1976), 216.

[3]Ibid., 215-221.

CHAPTER 5

CONCLUSIONS AND RECOMMENDATIONS

Purpose

This chapter provides the conclusions derived from the research methodology. It also includes recommendations concerning the Voice-Activated Logistics Anchor Desk and general recommendations concerning the military application of spoken language interface technology.

Conclusions

The purpose of this thesis is to answer the research question "Is a spoken language interface as part of the graphical user interface for the Logistics Anchor Desk more efficient than the graphical user interface alone?" Based primarily upon the system evaluation conducted as part of the research design of this project, it is the author's opinion that the combination of the graphical user interface with the spoken language interface is the most efficient configuration for the Logistics Anchor Desk. This configuration is known as the Voice-Activated Logistics Anchor Desk. Using the definition of efficiency provided in chapter 1, this determination is based on three criteria. The first is a comparison of the amount of time required to complete a task using the two different configurations of the Logistics Anchor Desk. The second criterion is subjective in that it requires a comparison of the amount of effort involved in using the two different configurations by each participant to complete the task scenario. This effort was measured comparing the number of steps a user had to complete and the number of times a user had to refer to the instruction sheet or seek assistance. The third and most critical criterion was the participant's impressions.

Time Criterion

The time data collected during the system evaluation strongly supports the conclusion that the spoken language interface increases efficiency. A comparison of the mean times for the sample population as well as each designated subgroup clearly shows that less time was required to complete the task scenario with the spoken language interface. The mean times for the sample population shows a decrease of sixty-six percent in time required with the interface. Group 1's times indicate a decrease of seventy-one percent. Those were the participants with the highest amount of computer experience. Group 2's times show a sixty six percent decrease in required time. Group three, those with the least amount of computer experience, showed a decreased time requirement of fifty-seven percent. Because of the small sample of this group, specific conclusions based upon a comparison of this percentage with the others cannot be made with any acceptable degree of confidence. Because this group's mean time falls within a standard deviation of the total sample population's time for this phase of the test, it is safe to say that they would follow the trend of time savings. A larger sample would be required to gain a more accurate analysis.

In addition to a comparison of the mean times of each group, another positive aspect of the results can be seen by analyzing the distribution of times for each configuration of the system. First of all, there is no overlap in the distributions of one standard deviation from each of the two means. This would indicate that the difference between the times is statistically significant and is not attributable to chance. Secondly, the distribution of times around the mean when the spoken language interface was used is much closer than when it was not used. It could be assumed from this that the spoken language interface is just as easy to use for the experienced user as it would be for a less experienced user. Because the time for each of the groups was decreased with the interface, it can be concluded that it provided an advantage for everyone that used the system and did not tend to favor one group over the other. This is an important point when considering the potential usability of the interface. It could be would concluded that the interface for the Voice-Activated Logistics Anchor

Desk would be more usable to a larger population of users than the interface for the Logistics Anchor Desk.

User Effort Criterion

Comparing the number of steps involved in completing a task with the two system configurations is straight forward. The number of steps required without the spoken language interface for each of the tasks in the scenario varied from four to ten. Each of those same tasks when using the spoken language interface requires only two steps, activate the microphone with the mouse and verbalize a command. An option that is available but was not activated for this test was the ability to activate the spoken language interface with a voice command, i.e., "Computer, ..." In this case, the only effort on the part of the user to complete a task with the computer is to speak. There can be no question that this is a more efficient technique than using the mouse and keyboard.

Comparing the number of times a user needed to refer to the instruction sheet with the number of times the observer had to assist participants with the spoken language interface clearly indicates less effort on the part of the participants when they use the spoken language interface. The type of assistance provided was different also. The instruction sheet was a detailed, step by step list of what the participant must do to complete the task, much like what would be found in a user's manual. In almost every case, the participants relied on the instructions in their entirety rather than using them as a memory jogger. When they required the observer's assistance during the second phase of the test with the spoken language interface, in all but one case the observer did not provide a detailed description of what needed to be done, but rather facilitated in their discovery of the solution. For example, when a participant used an out-of-vocabulary word repeatedly, the observer would suggest that they refer to the system dictionary rather than tell them which word the system would recognize. This technique was important for two reasons. First, it insured that the impressions they gained about the system included the use of its online help capabilities. Second, the average time it took them to complete the task scenario included the average amount of time a user at that experience level would take using the online help.

User Impression Criterion

Based upon the analysis of the data presented in chapter 4, it can be concluded that the participant's overall impression of the spoken language interface is that it is clearly more efficient than the current interface model. The data indicating the participant's overall impression and a statistically significant increase in the already positive impression after using the interface support this conclusion. Other aspects of the data that support this conclusion include the impressions of the utility of the interface and its ease of use. Written comments provided by the participants and observation records also support this conclusion. Another important impression the participants had about the interface is that it would reduce the training required for a user. This is an important aspect of the interface and would strongly support continued efforts by the Army to integrate it into the Voice-Activated Logistics Anchor Desk and many other automated systems.

General Conclusions

It has been the author's experience during the past sixteen years in the military that new automation systems fielded to units to increase productivity and reduce the time required to complete routine tasks are seldom useful if the leadership feels that the investment of time and people to complete the operator training is not worth the potential advantages that might be gained with the system. A good example of this phenomenon is what the author has seen occur with the first versions of the Standard Army Training System (SATS). This system is a software package that runs on a personal computer and automates most of the functions involved in the Army training management program. Functionally, the system can significantly reduce the amount of time it takes to complete tasks like producing weekly training schedules, conducting budget impact analysis on scheduled training events, building a unit's mission essential task list from digitized doctrinal products from the Army's training institutions, and so forth. The reality is that the user interface is so difficult to use that most military organizations the author has been associated with only invest the time it takes to learn how to produce a training schedule and for the most part don't use any of the other functions available. It seems very rare to find someone who is familiar with all the functions of the system or are even aware that more can be done than just produce a training schedule with it. A unit training manager designated to operate the system is required to attend a three-week course to operate the system and spends months getting comfortable with its operation. Since most soldiers spend no more than a year performing that duty, sometimes less, they have only begun to see the benefits of the system when it's time for them to assume other duties. The program is very functional but not very usable. The Army has developed a new version of the software that is more user friendly and can perform all of the previous functions and more. It would appear that the training requirement for this version remains about the same. The addition of a spoken language interface would reduce this requirement and make the program more usable for a greater number of people.

Another example based upon the author's personal experience demonstrates another aspect of this savings in training time. In a previous assignment the author was involved in the fielding of the Army's newest aerial signals intelligence collection and targeting system called Guardrail Common Sensor. The intelligence analyst workstation for that system is the same type of microcomputer used for the Voice-Activated Logistics Anchor Desk. It uses a graphical user interface with a keyboard and track ball to complete a multitude of analytical and intelligence reporting tasks. Operating the workstation requires a high degree of technical knowledge about data being collected by the system and the functions available on the workstation to manipulate that data. It also requires a high degree of skill on the part of the operator to maintain situational awareness while being exposed to a voluminous amount of information. The new equipment training schedule for the entire system was eight months. The training time required for one soldier to perform the basic functions of one of the workstations was a 252-hour curriculum. In this case, there was no question that the investment in time and people was necessary, the point is that an entire unit was not combat effective for nearly a year because of the training requirement associated with this system. This particular unit is one-of-a-kind in an Army corps and is its only organic source of aerial electronic intelligence on the battlefield. Reducing the training time in a case like this would have obvious benefits.

An additional benefit from the usability a spoken language interface would have in a case like this is important to note. The leadership involved with this system was well aware of the type of information that could be produced with it but were generally not familiar with the procedures required to produce it at a workstation because of the technical knowledge required. Having the ability to ask for a product through a spoken language interface instead of building it through data base searches and multiple electronic overlays would be an important advantage when time is of the essence.

Recommendations

The author recommends that the Army aggressively pursue the development of the Voice-Activated Logistics Anchor Desk. In addition, there should be efforts to insure spoken language interface capabilities are being included in all automated systems with graphical user interfaces currently under development. If the computer industry is proceeding on the assumption that speech is the next major component in the human-computer interaction model, it would follow that the Army should not be far behind.[1] For the Army to gain a decisive advantage in the new wave of warfare that changes in technology will bring, it must become a prominent player in the shaping and exploitation of that technology wave.[2] The basic tenets of the Army's Manpower and Personnel Integration (MANPRINT) process as a means of preparing itself for the challenges of warfare in the twenty-first century makes the exploitation of this technology necessary.[3] Any design issues and limitations that exist in this technology in its current state that impact military application will rapidly diminish if its development is closely tied to military requirements. "Whoever controls the development of human-machine interfaces may own the key to controlling information technology going into the next century."[4]

Endnotes

[1]Alexander I. Rudnicky, Alexander G. Hauptmann, and Kai-Fu Lee, "Survey of Current Speech Technology," Communications of the ACM 37 (March 1994): 52.

[2]Army Focus 1994, Force XXI, America's Army in the 21st Century (Washington: Department of the Army, 1994), 11.

[3]Ibid., 36-37.

[4]John Makhoul, Fred Jelinek, Larry Rabiner, Clifford Weinstein, and Victor Zue, Spoken Language Systems (Cambridge: BBN Systems and Technologies, 1990), 495.

APPENDIX A

INTERVIEW DATA

The following questions were the most common ones derived from the interviews as typical of the kinds of questions a logistics planner would need to answer:

1. What are the forecasted requirements for the units I am supporting at 24 hours, 48 hours, 72 hours, by commodity?

2. Where will a unit be in 24, 48, 72 hours?

3. What is the on-hand quantity of an item in a unit?

4. What is a unit's storage capacity?

5. What are the time/distance factors between this point and that point?

6. What is the Required Supply Rate for this unit? What is its Controlled Supply Rate and why?

7. Where are the ASPs, ATPs? Where are the units they support? What is their capability in short tons per day?

8. How many, by type, Combat Configured Loads are on hand?

9. What assets are available to move and handle supplies?

10. What roads are capable of handling movement of supplies?

11. How does terrain effect resupply?

12. Are there any railroads or pipelines available?

13. Where are the railheads?

14. Where are the units that require fuel other than/in addition to JP8?

15. What is the POL off load capability/storage capability of this port?

16. What ammo is on hand by weapons system and by DODIC? What is the condition code of that ammo?

APPENDIX B

QUESTIONNAIRES

Prestest Questionnaire

Student Number ____
Branch ___

1. How long have you been using computers?

_ < a year _ 1-5 years _ 5-10 years _ >10 years

2. How often do you use a desktop type computer? (Circle a number)

1	2	3	4	5
Daily	2 to 3 times per week	Once per week	A few times per month	Almost never

3. Which environment do you use most often when you operate a computer?

_ Windows _ DOS _ Macintosh _ Other

4. What is your experience with a voice interface used to operate a computer?

1	2	3	4
I use one regularly	I've tried one once or twice	I've seen one used	I've never seen or used one

5. How would you describe your impression of a voice interface capability to operate a computer?

1 2 3 4 5
|----------|----------|----------|----------|
Extremely positive — No opinion — Extremely negative

6. At what level or levels do you have logistics experience? (Check all that apply)

_ Brigade or below _ Division _ Corps _ EAC _ None

7. How would you describe your logistics skills?

1	2	3	4	5
Among the worst	Worse than most	About average	Better than most	Among the best

TASK SCENARIO

Student Number ______
Date ______

You have been tasked to review a TPFDD and make comments as a logistics planner. As part of this task, you must look at proposed locations for artillery units and Class V stockpiles at C+39 to determine effects of terrain and distance on resupply efforts. You will use the Logistics Anchor Desk (LAD) as a tool to conduct your analysis. The TPFDD you are using is part of a corps contingency plan for operations in South Korea.

To begin your analysis, you must execute five separate tasks on the LAD to display the information you need. To evaluate the performance of the LAD with a voice interface, you will execute the tasks without the voice interface first, then with the voice interface. The tasks you must execute are listed below in sequence. Please note the information requested as you work through the sequence of tasks. You will complete a post-test questionnaire to record your comments and impressions of the voice interface capability for the LAD.

1. Record the time now to the nearest second. (i.e., 1053 hrs 37 sec.)
   _____ hrs ___ sec

2. Display the area encompassing South Korea. If you are not sure how to complete this task with the mouse, refer to the instruction sheet titled "Display South Korea" provided in the manilla folder. If you have to refer to the instructions, place a check here__.

3. Perform a query by NSN and display the available Class V on the map of South Korea. If you are not sure how to complete this task with the mouse, refer to the instruction sheet titled "Display Class V" provided in the manilla folder. If you have to refer to the instructions, place a check here__.

4. Perform a query by unit and display all the FA units on the map of South Korea. If you are not sure how to complete this task with the mouse, refer to the instruction sheet titled "Display FA Units" provided in the manilla folder. If you have to refer to the instructions, place a check here__.

5. Place a box around the area covered by the Class V stockpiles and units. To do this with the mouse, move the pointer to the upper left corner of the area, push and hold the left mouse button and drag the pointer to the lower right corner of the area to form the box around all the units and Class V stockpiles.

6. Zoom in on the boxed-in area. If you are not sure how to complete this task with the mouse, refer to the instruction sheet titled "Zoom In on Selected Area" provided in the manilla folder. If you have to refer to the instructions, place a check here__.

7. Overlay the JOG-A on the display. If you are not sure how to complete this task with the mouse, refer to the instruction sheet titled "Overlay a JOG-A" provided in the manilla folder. If you have to refer to the instructions, place a check here__.

8. Record the time now to the nearest second. (i.e., 1053 hrs 37 sec.)
   _____ hrs ___ sec

9. The observer will give you a brief demonstration of the voice interface before you go on to the next phase.

10. Record the time now to the nearest second. (i.e., 1053 hrs 37 sec.)
____ hrs ___ sec

11. Display the map of South Korea.

12. Display Class V on the map display.

13. Display the field artillery units on the map display.

14. Place a box around the area covered by the Class V stockpiles and units. Do this with the mouse by moving the pointer to the upper left corner of the area, push and hold the left mouse button and drag the pointer to the lower right corner of the area to form the box around all the units and Class V stockpiles. When that is completed, ask the computer to "go there."

15. Overlay the JOG-A on the map display.

16. Record the time now to the nearest second. (i.e., 1053 hrs 37 sec.)
____ hrs ___ sec

PostTest Questionnaire

Student Number ____

1. How do you feel about the amount of time it took to complete the scenario with the voice interface? (Circle the number)

1	2	3	4	5
Much quicker than I thought	Quicker than I thought	About right	Longer than I thought	Much longer than I thought

2. How would you complete this sentence: The amount of time it took to complete the tasks in this scenario with the voice interface is _____. (Circle the letter)

a. as fast as I would like for actual work
b. fast enough for most work
c. tolerable for some work
d. too slow for most work
e. too slow for all work

3. Using the voice interface was:

1	2	3	4	5
Very easy	Easy	About right	Difficult	Very difficult

4. How would you describe your impression of the voice interface?

1	2	3	4	5
Very impressed	Impressed	No opinion	Disappointed	Very disappointed

Please explain why you have that impression.

____

____

____

5. Understanding that the voice interface you used is currently under development, How would you describe your recommendation concerning the Army's continued effort to integrate a voice interface into the LAD?

1	2	3	4	5
Highly positive	Positive	Neither positive or negative	Negative	Highly negative

Explain why you would make that recommendation.

6. Please make any additional comments about this voice interface or voice interface capabilities in general for military application.

APPENDIX C

TASK INSTRUCTIONS

Display South Korea

1. With the mouse, move the pointer to the small globe icon with the word "Map" on it in the upper left portion of the LAD display and click the left mouse button.

2. Move the pointer to the words "Go to" below and to the left of the globe and click the left mouse button.

3. Move the pointer down to the words "Named Views", a second menu will appear to the right. Move the pointer strait across to that menu.

4. Move the pointer down to the words "S. Korea" and click the left mouse button.

Display Class V

1. Move the pointer to the word "NSNs" in the upper right portion of the LAD display and click the left mouse button.

2. Move the pointer to the words "CLASS V" on the left side of the display and click the left mouse button.

3. Move the pointer to the globe icon used earlier and click the left mouse button.

Display FA Units

1. Move the pointer to the word "Units" below "NSNs" and click the left mouse button.

2. Move the pointer to the word "Query" on the left side of the display and click the left mouse button.

3. Move the pointer to the words "Troop_List" below "Query" and click the left mouse button. The words "Troop_List" will appear below the first one. Move the pointer to that "Troop_List" and click the left mouse button. A menu will appear.

4. Because of a software bug, the bottom half of that menu is cut off the first time you activate it. For that reason, you must repeat step #3 to see the entire menu.

5. On the menu that appears, move the pointer to the bottom on the word "<<<more>>>", a second menu will appear to the right. Move the pointer strait across to that menu, go up to the word "Nomenclature" then click the left mouse button.

6. On the menu that appears, move the pointer to the word "Like" and click the left mouse button.

7. A blinking cursor will appear in a small window just above the area where the menu was displayed. Type in the letters "fa" and hit the "Enter" key.

8. Move the pointer to the "Map" icon and click the left mouse button.

<u>Zoom In on Selected Area</u>

1. Move the pointer to the words "Go To" on the left side of the display and click the left mouse button.

2. On the menu that appears, move the pointer to the words "Selected Area" and click the left mouse button.

<u>Overlay a JOG-A</u>

1. Move the pointer to the word "Layers" near the center of the LAD display and click the left mouse button.

2. From the menu that appears, move the pointer to the word "Edit" and click the left mouse button.

3. A small right angle with an arrow will appear. Move this image to the upper portion of the screen off of the map area and click the left mouse button. A small image like a spreadsheet table will appear, click the left mouse button again.

4. On the menu that appears will be two small windows. Scroll through the list of choices in the lower window by moving the pointer to the small down arrow to the right of the lower window and holding the left mouse button down. Scroll the choices until you see "ADRG-JOG-A". Move the pointer to that choice and double click the left mouse button. You should see that choice appear in the upper window.

5. Move the pointer to the word "OK" at the bottom of the menu and click the left mouse button.

APPENDIX D

VALAD USABILITY TEST DATA

Table 1--VALAD usability test data

Participant	General computer experience	Desktop use	Primarily GUI	Exposure to SLS	Impression of SLS	Logistic experience	Proficiency	Time without SLS	Needed instructions
3752	2	2	YES	4	3	1	4		
3765	2	1	YES	4	2	1,3	3		
3825	2	1	YES	3	3	1,2	4		
3856	2	1	YES	4	4	1-3	4		
3914	2	2	NO	3	3	1,2	3		
4007	2	2	YES	4	3	1	3	0:10:23	100%
4500	2	1	YES	4	3	1,4	3		
4588	2	5	YES	3	1	1	5		
4651	2	3	YES	4	2	1	4	0:12:00	100%
4682	2	3	YES	3	1	1	4	0:08:10	100%
4753	2	2	YES	3	2	5	1		
Inst.#2	3	1	YES	2	2	1,3	4	0:09:18	20%
3298	3	2	YES	4	2	1	3		
3850	3	2	YES	4	3	1-3	4		
3902	3	2	YES	3	2	1,4	4	0:07:18	100%
4016	3	2	YES	4	3	3,4	3		
4079	3	1	YES	4	2	1-3	3	0:11:51	100%
4138	3	1	YES	4	3	1	3		
4150	3	2	YES	2	2	1,4	3		
4230	3	2	YES	3	2	2	4		
4239	3	1	YES	4	3	2,3	4	0:10:53	100%
4277	3	2	YES	3	1	1	3	0:16:05	100%
4366	3	2	YES	4	3	1,4	4		
4508	3	1	YES	2	2	2	3	0:08:00	100%
4630	3	1	YES	4	3	1,2	3	0:09:20	100%
4632	3	1	YES	4	2	1	3		

Table 1--Continued

Participant	General computer experience	Desktop use	Primarily GUI	Exposure to SLS	Impression of SLS	Logistic experience	Proficiency	Time without SLS	Needed instructions
4737	3	1	YES	4	3	1-4	3.5		
4758	3	1	YES	4	4	1,4	3	0:07:17	100%
4782	3	1	NO	4	2	1,2	5		
*4794	3	1	YES	4	3	4	4	0:08:48	100%
4825	3	2	YES	4	2	1-3	4		
4856	3	1	NO	3	2	1-3	4	0:12:10	100%
Inst.#1	4	1	YES	2	2	1-4	5	0:08:18	20%
3744	4	1	YES	4	3	1	3		
3822	4	1	YES	4	1	1,2,4	5		
4059	4	1	YES	4	1	1,2	4		
4065	4	2	YES	4	3	1	3		
4164	4	1	YES	2	1	1	3	0:12:35	60%
4179	4	2	YES	3	3	1	3		
4183	4	4	YES	4	2	1	3	0:07:58	100%
4219	4	1	YES	3	2	1-3	3	0:18:40	100%
4255	4	1	YES	1	2	1,2	5		
4272	4	2	YES	4	3	3,4	3		
4367	4	1	NO	4	3	1,2,4	5	0:09:00	100%
*4414	4	1	YES	4	2	1,3,4	3	0:04:05	100%
4449	4	1	YES	2	3	1,4	3	0:07:19	100%
4452	4	5	YES	3	3	1	4		
4467	4	1	YES	2	2	1	3		
4517	4	1	YES	3	2	1-3	3		
4679	4	1	YES	4	3	2-4	4	0:06:30	100%
4685	4	1	YES	2	2	1	3		
4714	4	1	YES	3	2	1	3.5		

Table 1--Continued

Participant	General computer experience	Desktop use	Primarily GUI	Exposure to SLS	Impression of SLS	Logistic experience	Proficiency	Time without SLS	Needed instructions
4799	4	2	YES	2	2	1	3		
Inst.#3	4	1	YES	2	2	1,2,5	5	0:04:00	20%
Averages	3.15	1.56	50 YES,4 NO	3.25	2.31		3.49	0:08:34	75%
Std. dev.	0.86	0.95		0.94	0.78		0.91	0:04:32	
Group 1 averages	4.00	1.50		3.00	2.23		3.61	0:08:16	67%
Group 2 averages	3.00	1.43		3.24	2.31		3.51	0:09:13	82%
Group 3 averages	2.00	2.09		3.24	2.30		3.50	0:10:11	100%
Instructors	3.67	1		2	2		4.67	0:07:12	20%
Participants	3.25	1.5		3.15	2.25	(Std. dev.=0.77)	3.65		

Table 1--Continued

Participant	Time with SLS	Needed assistance	Impression of usability	Estimate of utility	Ease of use	Overall impression	Recommend-ation
3752							
3765							
3825							
3856							
3914							
4007	0:03:50	NO	2	2	3	2	2
4500							
4588							
4651	0:05:03	YES	4	1	2	1	1
4682	0:04:15	NO	2	1	1	1	1
4753							
Inst.#2	0:02:00	NO	3	1	1	1	1
3298							
3850							
3902	0:05:28	NO	2	3	2	2	1
4016							
4079	0:02:31	NO	1	2	1	1	1
4138							
4150							
4230							
4239	0:05:34	NO	1	2	1	1	1
4277	0:03:01	NO	1	1	1	1	1
4366							
4508	0:03:00	YES	1	2	2	2	2
4630	0:03:52	NO	1	2	1	1	1
4632							

Table 1--Continued

Participant	Time with SLS	Needed assistance	Impression of usability	Estimate of utility	Ease of use	Overall impression	Recommend-ation
4737							
4758	0:01:49	NO	1	3	1	1	1
4782							
*4794	0:01:28	NO	1	NA	1	1	1
4825							
4856	0:04:00	NO	1	2	1	1	1
Inst.#1	0:02:20	NO	2	2	1	1	1
3744							
3822							
4059							
4065							
4164	0:03:55	NO	2	1	1	1	1
4179							
4183	0:05:20	NO	1	1	1	2	2
4219	0:02:25	YES	2	2	2	1	1
4255							
4272							
4367	0:02:55	YES	1	1	1	1	1
*4414	0:06:50	YES	4	3	2	4	2
4449	0:01:31	NO	1	1	2	2	2
4452							
4467							
4517							
4679	0:01:50	NO	2	2	1	2	2
4685							
4714							

Table 1--Continued

Participant	Time with SLS	Needed assistance	Impression of usability	Estimate of utility	Ease of use	Overall impression	Recommend-ation
4799							
Inst.#3	0:01:35	NO	2	2	2	2	2
Averages	0:02:53		1.43	1.48	1.22	1.17	1.13
Std. dev.	0:01:39		0.92	0.83	0.72	0.64	0.61
Group 1 averages	0:02:26		1.44	1.33	1.22	1.33	1.33
Group 2 averages	0:03:08		1.20	1.80	1.10	1.10	1.00
Group 3 averages	0:04:23		2.67	1.33	2.00	1.33	1.33
Instructors Participants	0:01:58		2.33	1.67	1.33	1.33	1.33

BIBLIOGRAPHY

Articles

Allen, James F., George Ferguson, Brad Miller, and Eric Ringger. Spoken Dialogue and Interactive Planning. Rochester: Department of Computer Science, University of Rochester, 1995.

Allen, James F., Lenhart K. Schubert, George Ferguson, Peter Heeman, Chung Hee Hwang, Tsuneaki Kato, Marc Light, Nathaniel G. Martin, Bradford W. Miller, Massimo Poesio, and David R. Truman. "TheTRAINS Project: A Case Study in Building a Conversational Planning Agent." Journal of Experimental and Theoretical AI 7 (1995): 7-48.

Hawkins, Donald T. "Breaking the Keyboard Barrier: Voice Input To Information Retrieval Systems." ONLINE, November/December 1994, 66.

Isaacson, Portia. "Wildfire's Telephony Assistant May Be Your Best Friend." Computer Reseller News, 1 May 1995, 63.

Jones, Karen Spark. "Towards Better NLP System Evaluation." In ARPA Workshop on Human Language Technology, March 8-11,1994. Washington: Advanced Research Projects Agency Information Science and Technology Office, 1994, 98-103.

Krivda, Cheryl D. "GUI Takes the Task." Midrange Systems 8 (14 April 1995): 28-30.

Orr, Joel N. "In Pursuit of Immediacy." Computer-Aided Engineering 14 (February 1995): 66.

Rudnicky, Alexander I., Alexander G. Hauptmann, and Kai-Fu Lee. "Survey of Current Speech Technology." Communications of the ACM 37 (March 1994): 52-57.

Zarowin, Stanley. "Touchless Typing." Journal of Accountancy 179 (April 1995): 63-65.

Compton's Interactive Encyclopedia, 1994 ed. S.v. "Typewriter."

Books

Dumas, Joseph S. and Janice C. Redish. A Practical Guide to Usability Testing. Nolwood, NJ: Ablex Publishing Corporation, 1993.

Herzon, Frederick D., and Michael Hooper. Introduction to Statistics for the Social Sciences. New York: Harper & Row, Publishers, Inc., 1976.

Technical Reports and Reviews

Andes, William S., Timothy Barry, David Hartsock, John A. Kovacs, Peter B. Lovering, James R. McCracken, Michael C. Reynolds, and Russel A. Temperley. Pilot Factors Research for Crew Systems Development. Dayton, OH: Midwest Systems Research, Inc., 1993.

Bates, Madeline, Robert Bobrow, Francis Kubala, Robert Ingria, John Makhoul, Scott Miller, Long Nguyen, Sandra Peters, Richard Schwartz, David Stallard, and George Zavaliagkos. Usable, Real-Time, Interactive Spoken Language Systems. Cambridge: BBN Systems and Technologies, 1994.

Carroll, Edward J. and Ken Lam. Voice/Natural Language Interfacing for Robotic Control. Dover, NJ: Army Armament Research Development and Engineering Center, 1987.

Cooke, N. RAE Bedford's Experience of Using Direct Voice Input (DVI) in the Cockpit. United Kingdom: Royal Aerospace Establishment, Farnborough, 1990.

Flater, D. and Y. Yesha. Coping with Different Retrieval Methods in Next Generation Networks. Gaithersburg, MD: National Institute of Standards and Technology, Information Systems Engineering Division, 1994.

Galliers, J. R. and Karen Sparck Jones. Evaluating Natural Language Processing Systems. Cambridge, UK: Computer Laboratory, University of Cambridge, 1993. TR 291.

Kersteen, Z. A. and D. Damos. Human Factors Issues Associated with the Use of Speech Technology in the Cockpit. Tempe, AZ: Arizona State University, Department of Psychology, 1989.

Makhoul, John, Fred Jelinek, Larry Rabiner, Clifford Weinstein, and Victor Zue. Spoken Language Systems. Cambridge, MA: BBN Systems and Technologies Corp., 1990

Makhoul, John, and Richard Schwartz. State of the Art in Continuous Speech Recognition. Cambridge, MA: BBN Systems Miller, D. P. Top Ten List of User-hostile Interface Design. Albuquerque, NM: Sandia National Labs, 1994.

Plutchak, B. D. Voice Operated Status Boards for the CATCC (Carrier Air Traffic Control System). San Diego, CA: Naval Oceans Systems Center, 1988.

Stokes, James M. Voice Data Entry in NISTARS Warehouses. Spring House, PA: CHI Systems, Inc., 1993.

Weinstein, C. J. Opportunities for Advanced Speech Processing in Military Computer-based Systems. Boston, MA: Massachusetts Institute of Technology, Lexington Lincoln Lab, 1991.

Annual Research Summary, Spoken Language Systems, January 1995. Cambridge, MA: Laboratory for Computer Science, Massachusetts Institute of Technology, 1995.

Internet Source Material

Polson, Ken. (1995. November 4). Chronology of Events in the History of Microcomputers [On-line]. Available: http://www.islandnet.com/~kpolson/comhist.htm

(1995, November 4). Human-Computer Interaction Program [On-line]. Available: http://www.csto.arpa.mil/ResearchAreas/Intelligent_Systems/Human_Computer_Interaction.html

(1995, November 4). Human Language Systems Program [On-line]. Available: http://www.csto.arpa.mil/Research/Intelligent_Systems/Human_Language_Systems.html

Interviews

Atkinson, Gloria, Lieutenant Colonel, Quartermaster, United States Army. Interview by author, 16 November 1996, Fort Leavenworth, KS. Personal notes.

Bochsler, Hal, Major, Quartermaster, United States Army. Interview by author, 15 November 1996, Fort Leavenworth, KS. Personal notes.

Eisenbeis, Jeffrey A., Lieutenant Colonel, Ordnance, United States Army. Interview by author, 30 November 1996, Fort Leavenworth, KS. Personal notes.

Unpublished Materials

Carthon, Michel. The Telecommunications Emergency Decision Support System as a Crisis Management Decision Support System. Monterey, CA: Naval Postgraduate School, 1991.

Hill, Jerry K. Application of Speech Recognition to the Integrated Tactical Decision Aid (ITDA). Monterey, CA: Naval Postgraduate School, 1988.

Hoeller, Mary L. and Claudia R. Salerni. Using Continuous Voice Recognition to Operate the Research Evaluation System Analysis (RESA) Wargame. Monterey, CA: Naval Postgraduate School, 1990.

Leventis, Constantinos P. Speech Recognition Application in C.I.C. Monterey, CA: Naval Postgraduate School, 1991.

Powers, Richard A. Continuous Speech Recognition as an Input Method for Tactical Command Entry in the SH-60B Helicopter. Monterey, CA: Naval Postgraduate School, 1992.

Scardefield, Stephen S. Analysis of Manual Versus Voice Response Technology for System Interface Design. Monterey, CA: Naval Postgraduate School, 1992.

Government Publications

U.S., Department of the Army. Army Focus 1994, Force XXI: America's Army In The 21st Century. Washington, DC: Government Printing Office, 1994.

U.S., Department of the Army. US Army Command and General Staff College. ST 101-5, Command and Staff Decision Processes. Fort Leavenworth: USACGSC, February 1995.

INITIAL DISTRIBUTION LIST

1. Combined Arms Research Library
   U.S. Army Command and General Staff College
   1 Reynolds Ave.
   Fort Leavenworth, KS 66027-1352

2. Defense Technical Information Center
   Cameron Station
   Alexandria, VA 22314

3. LTC James C. McNaughton
   5344 Soquel Drive
   Soquel, CA 95073

4. LTC Jerry L. Thompson
   DLRO
   USACGSC
   1 Reynolds Ave.
   Fort Leavenworth, KS 66027-1352

5. MAJ (P) John S. Kane
   DLRO
   USACGSC
   1 Reynolds Ave.
   Fort Leavenworth, KS 66027-1352

6. Gary K. Poock
   Naval Postgraduate School
   Monterey, CA 93943-5000

7. Madeleine Bates
   c/o BBN Systems and Technologies
   70 Fawcett Street
   Cambridge, MA 02138

CERTIFICATION FOR MMAS DISTRIBUTION STATEMENT

1. Certification Date: 15 / 04 / 96

2. Thesis Author: MAJ Theron Bowman

3. Thesis Title: An Evaluation of the Spoken Language System Interface for the Voice-Activated Logistics Anchor Desk

4. Thesis Committee Members Signatures:

5. Distribution Statement: See distribution statements A-X on reverse, then circle appropriate distribution statement letter code below:

A B C D E F X SEE EXPLANATION OF CODES ON REVERSE

If your thesis does not fit into any of the above categories or is classified, you must coordinate with the classified section at CARL.

6. Justification: Justification is required for any distribution other than described in Distribution Statement A. All or part of a thesis may justify distribution limitation. See limitation justification statements 1-10 on reverse, then list, below, the statement(s) that applies (apply) to your thesis and corresponding chapters/sections and pages. Follow sample format shown below:

S------------SAMPLE----------SAMPLE-----------SAMPLE-----------SAMPLE-----------S

	Limitation Justification Statement	Chapter/Section	Page(s)	
A				A
M				M
P	Direct Military Support (10)	Chapter 3	12	P
L	Critical Technology (3)	Sect. 4	31	L
E	Administrative Operational Use (7)	Chapter 2	13-32	E

-------------SAMPLE----------SAMPLE-----------SAMPLE----------SAMPLE------------

Fill in limitation justification for your thesis below:

Limitation Justification Statement	Chapter/Section	Page(s)

7. MMAS Thesis Author's Signature: